Lester S. Kellogg

Library

St. Olaf College

Northfield, Minn.

Digitized by the Internet Archive
in 2026

https://archive.org/details/forecastingyield0000henr

FORECASTING THE YIELD AND THE PRICE OF COTTON

THE MACMILLAN COMPANY
NEW YORK · BOSTON · CHICAGO · DALLAS
ATLANTA · SAN FRANCISCO

MACMILLAN & CO., LIMITED
LONDON · BOMBAY · CALCUTTA
MELBOURNE

THE MACMILLAN CO. OF CANADA, LTD.
TORONTO

FORECASTING THE YIELD AND THE PRICE OF COTTON

BY

HENRY LUDWELL MOORE

PROFESSOR OF POLITICAL ECONOMY IN COLUMBIA UNIVERSITY
AUTHOR OF "ECONOMIC CYCLES: THEIR LAW AND CAUSE," AND OF "LAWS OF WAGES"

"We have to contemplate social phenomena as susceptible of prevision, like all other classes, within the limits of exactness compatible with their higher complexity."

AUGUSTE COMTE.

THE NATIONAL SHAWMUT BANK OF BOSTON

New York
THE MACMILLAN COMPANY
1917

All rights reserved

HD9074
.M7

Copyright, 1917
By THE MACMILLAN COMPANY
Set up and printed. Published October, 1917.

NAT. SHAWMUT BANK
G

CONTENTS

CHAPTER I

PAGE

Introduction 1

CHAPTER II

THE MATHEMATICS OF CORRELATION

A Frequency Distribution 17
The Standard Deviation as a Measure of Dispersion . . 20
The Fitting of Straight Lines to Data 28
The Coefficient of Correlation 40

CHAPTER III

THE GOVERNMENT CROP REPORTS

The Character and the Aim of the Crop-Reporting Service 52
Technical Terms: Normal, Condition, Indicated Yield per Acre 58
The Accuracy of Forecasts Tested 65
Acreage and Production 82

CHAPTER IV

FORECASTING THE YIELD OF COTTON FROM WEATHER REPORTS

The Official Forecasts of the Yield of Representative States 94
Forecasting the Yield of Cotton from the Accumulated Effects of the Weather 100
The Results Compared for the Representative States . . 115
Three Possible Objections 121

CHAPTER V

THE LAW OF DEMAND FOR COTTON

PAGE

Two Practical Methods of Approach 140
Statics and Dynamics Discriminated 147
A Complete Solution of the Problem 151

CHAPTER VI

Conclusions 163

FORECASTING THE YIELD AND THE PRICE OF COTTON

FORECASTING THE YIELD AND THE PRICE OF COTTON

CHAPTER I

INTRODUCTION

An eminent economist has recently told us that economists no longer talk so confidently as they once did of forecasting social phenomena, and that, confronted with the complexity of social relations, "the sober-minded investigator will be slow in laying too much stress on single causes, slow in generalization, slowest of all in prediction." An equally distinguished statistician has warned his colleagues of the dangers of using refined mathematical methods in the treatment of the loose data supplied by our official bureaus. These are authoritative warnings, and I have not been unmindful of them as the successive theses of this Essay have been developed. But the ultimate aim of all science is prediction; the most ample and trustworthy data of economic science are official statistics; and the only adequate means of exploiting raw statistics are mathematical methods.

The statistical devices used in the treatment of our problem of forecasting prices were, for the most part, invented, for another purpose, by Professor Karl Pearson, and rest upon the theory of probabilities. Of the

Pearsonian superstructure one may repeat what Laplace has said of its foundation: "*que la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul; elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct, sans qu'ils puissent souvent s'en rendre compte.*" On the Cotton Exchanges of the world there are always certain speculators, *les esprits justes* of the commodity market, who seem to know by a kind of instinct the degrees of significance to attach to Government crop reports, weather reports, changes in supply and demand, and the movements of general prices. Mathematical methods of probability reduce to system the extraction of truth contained in official statistics and enable the informed trader to compute, with relative exactitude, the influence upon prices of routine market factors.

The Department of Agriculture of the United States, referring to the use of its crop-reporting service, has briefly described the aim of its admirable statistical organization:

"Everything . . . which tends toward certainty, as regards either supply or demand, is distinctly advantageous to the farmer. Hence to throw light on future conditions and do away, as far as possible, with uncertainties as to supply and demand, is the principal object of the statistical work of the Department of Agriculture and constitutes the sole reason for the collection of data and the publication of information regarding current accumulations of farm products and concerning crop conditions and prospects. In so far, therefore, as these data are accurate and reliable —

qualities which depend on the integrity and intelligence of crop correspondents and their interest in the work — the publication of the information secured can not fail to reduce the uncertainty regarding the future values of farm products, and thus have an important cash value to all farmers."

Without a doubt great values are at stake. If the size of the cotton crop of 1914 is taken as a standard, an error in an official crop report which should lead to an ultimate depression of one cent a pound in the price of cotton lint would cost the farmers $80,000,000, or more. A corresponding error leading to a similar rise in price would entail upon manufacturers and consumers a comparably heavy loss.

The Department of Agriculture has rendered its reports continuously for some fifty years, and yet, as far as I am aware, no one has either measured the degree of accuracy of the information it supplies "concerning crop conditions and prospects," or attempted to see whether, by different methods, more truth might not be gained from the stores of raw figures that its Bureaus collect.

Government Departments seeking appropriations are very likely, out of administrative necessity, to stress their successes and suppress their failures. In the January number of the official *Crop Reporter*, for 1900, this illustration is given of the value to planters of the Government crop-reporting service:

"The past year has afforded a striking example of the influence of the reports on prices. As early as August the Division of Statistics called attention to the prevailing drought and its deleterious effect upon the

growing crops. July 1 the average condition reported was 87.8; August 1 it was 84; September 1 it was 68.5; October 1, 62.4; resulting in an average estimated yield of lint cotton per acre of 184 pounds. Now, the lowest price of futures in the New York Cotton Exchange during 1899 was reached June 29 when July deliveries sold at 5.43. The highest price was November 9, when July deliveries sold at 7.74. Commercial authorities of high standing had strongly disputed the position taken by the Division of Statistics, their estimates running as high in some cases as 12,000,000 bales. To-day the Department estimate of December 10 of 8,900,000 bales is generally conceded to be very close to the truth, even by these same commercial authorities. While, therefore, the effect of these overestimates was only temporary, it was, nevertheless, sufficient to cause a loss of several millions to the cotton planters."

This is a success defiantly stressed. We shall have to take but a step to come upon a failure ingloriously suppressed. The reference to the preceding instance is given in the January number of the *Crop Reporter*, for 1900, p. 2. But in this same year 1900, the *Crop Reporter* for July, p. 2, gives the following account of the condition and prospects of the cotton crop for the current year:

"Not only was the condition on July 1 for the cotton region as a whole the lowest July condition on record, but in Georgia, Florida, Alabama, and Mississippi also it was the lowest in the entire period of 34 years for which records are available, while in Tennessee it was the lowest with one exception and in South Carolina, Texas, and Arkansas the lowest with two excep-

tions in the same period of 34 years. Excessive rains, drowning out the crop, and followed by an extraordinary growth of grass and weeds, are reported for almost every State, and the gravity of the situation is greatly increased by the general scarcity of labor. In South Carolina, Georgia, Alabama, Louisiana, and Texas considerable areas will have to be abandoned."

Notwithstanding this ill-boding forecast, the records of the Bureau of Statistics show that the yield per acre for 1900 was, with the exception of two years, the largest in three decades. Later on, as the crop approached maturity, the successive monthly reports departed more and more from the early forecast and then the official Bureau issued a final estimate that approximated the truth. When, however, Secretary Wilson of the Department of Agriculture was seeking with the help of Senator W. B. Allison to prevent the duplication by the Census Bureau of work usually done by his Department, he refers to the final estimate, by his Department, of the cotton crop of this same year, 1900–1901, as "an estimate so accurate that its subsequently ascertained close agreement with actual production was commented upon throughout the entire cotton world as a marvel of statistical forecasting." (*Crop Reporter*, March, 1902, p. 4.)

Now, obviously, what is needed for business and for scientific purposes is not one or more illustrations either in praise or in blame of the Government crop-reporting service, but a quantitative testing of the accuracy of the continuous service throughout a long period, say a quarter of a century. For business and for scientific purposes one must know the degree of accuracy with

which, upon an average, from the official data available at any time, one can forecast the ultimate yield.

On many, if not upon all the Cotton Exchanges of the country, the daily variations of rainfall and temperature in the states of the Cotton Belt, during the growing season, are, for the information of brokers, plotted on a large map. The Government crop reports describe the variations of weather during the interval covered in their crop survey. The leading newspapers give daily reports of the "Weather in the Cotton Belt." To-day, August 18, 1916, the *New York Times* prints a typical description of the influence of weather reports on trading and prices:

COTTON ADVANCES IN STEADY MARKET

STORM IN THE GULF OF MEXICO KEEPS THE TALENT GUESSING

BUT RAINS HELP TEXAS CROP

"There was a steady undertone in the cotton market yesterday, but trading was rather light when the wide-spread interest in cotton at this season is taken into consideration. There was a manifest disposition to wait for further crop developments, and the fact that there was a storm in the Gulf of Mexico working toward the cotton belt made both the longs and the shorts a bit timid. On one hand there was the possibility that this disturbance might bring much needed rain to the region west of the Mississippi; on the other hand the danger that it might give bad weather to the Southern Atlantic States, where there has been severe damage by storms and too much rain. . . .

"Private reports from the belt were not particularly bullish, as they told of scattered showers in Texas and improvement in some parts of the Eastern States. The talk of the approaching storm, however, rather overshadowed the reports of the weather of the

minute, although the bulls did not neglect to call attention to the fact that there was no relief in Oklahoma, where rain was much needed."

A series of critical questions is suggested by the great importance which Government Bureaus, Cotton Exchanges, and the Public Press very obviously attach to the weather conditions as they are related to the cotton crop:

(1) Variations of both temperature and rainfall must affect the yield per acre of cotton, but do they affect the yield in the same way and to the same degree? What is the measure of the effect of each, independent of the other? What is the measure of their joint effect?

(2) In Texas, throughout a quarter of a century, the yield of cotton has been steadily falling, while in Georgia, throughout the same interval, the yield has been steadily increasing. How, then, can one measure the effects, jointly and separately, of rainfall and temperature upon the crop of each state and upon their combined crop?

(3) Suppose that the above questions are satisfactorily answered for one particular month. Would the answers be different for different months? Or would the particular combination of temperature and rainfall for, say, July, produce an equal effect with the same combination for August? Are the answers for this and the above questions the same for all of the cotton-producing states?

(4) Supposing that one has solved the above ques-

tions for all of the states of the Cotton Belt, how could one take account of the variations of the weather from the beginning of the growing season up to a given date, in such a way as to be able to forecast their possible joint effects upon the ultimate yield of cotton?

(5) Supposing that one could forecast the yield per acre of cotton from the successive reports of the Weather Bureau, how would the degree of accuracy of such forecasts compare with the forecasts of the Crop-Reporting Board, which are based upon the direct observations of the thousands of correspondents of the Department of Agriculture?

A knowledge of the acreage and of the probable yield per acre of cotton will afford the necessary data to compute the probable supply. But in order to forecast the probable price of cotton lint, the law of demand for cotton must be known. That is to say, one must know the probable variation in the price that will accompany a computed variation in the supply.

With regard to this question of demand economic science is in the state which electrical science had reached about the middle of the nineteenth century. It would appear that there are two sciences of economics, one of the class room and one of the market place, and the difference between the two is the same as the difference described by Fleeming Jenkin as existing between the Electricity of the Schools and the Electricity of the Practical Engineer:

"The difference between the Electricity of Schools

and of the testing office has been mainly brought about by the absolute necessity in practice for definite measurement. The lecturer is content to say, under such and such circumstances, a current flows or a resistance is increased. The practical electrician must know how much current and how much resistance, or he knows nothing."

The *Open Sesame* to academic economics is the "law of supply and demand" or "the equation of demand and supply." No general problem within the confines of the science may be approached except through the "law of supply and demand." But, as incredible as it may seem, what the law of demand actually is for any one commodity is nowhere stated in the text-books. Indeed not only do the text-book writers forbear to state the law for any one commodity, but, as a rule, they either omit to say whether there is any hope of ever knowing the law in any concrete case, or else say bluntly that the law can never be known because their discussion of economic theory is confined to normalities within an hypothetical, static state. The economist of the market place, however, not only must know that, under given circumstances of the supply, the price will rise or fall, but he must know the probable limits within which the price fluctuations will be confined.

In different ways many agencies, public and private, assemble facts that have a bearing upon the probable demand for cotton, and the findings of the several inquiries are published for the information of those directly or indirectly concerned. Each individual is left free to draw his own conclusions as to the joint effect of the many factors in the problem, and the re-

sulting conduct of the many buyers gives definiteness to the law of demand for cotton. Would it not be possible to describe this resulting law of demand with a degree of precision as great as the accuracy with which, from elaborate Government reports as to crop-conditions and crop-prospects, the official Bureau forecasts the probable supply of cotton?

By means of the principles and methods presently to be described, it is possible for any person (1) from the current reports of the Weather Bureau as to rainfall and temperature in the states of the Cotton Belt, to forecast the yield of cotton with a greater degree of accuracy than the forecasts of the Department of Agriculture, and (2) from the prospective magnitude of the crop, to forecast the probable price per pound of cotton with a greater precision than the Department of Agriculture forecasts the yield of the crop.

The principal purpose of my Essay I should like to make very clear. It is not to point out the limitations of the work done in forecasting by the Department of Agriculture; much less, to urge any device of my own as a substitute for the methods that are followed by the official Statistical Bureaus. My chief aim has been to make a contribution to economic science by showing that the changes in the great basic industry of the South which dominate the whole economic life of the Cotton Belt are so much a matter of routine that, with a high degree of accuracy, they admit of being predicted from natural causes.

The business of economic science, as distinguished

from economic practice, is to discover the routine in economic affairs. It aims to separate out the elements of the routine, to ascertain their interdependence, and to use the knowledge of their connections to anticipate experience by forecasting from known changes the probabilities of correlated changes. The seal of the true science is the confirmation of the forecasts; its value is measured by the control it enables us to exercise over ourselves and our environment.

CHAPTER II

THE MATHEMATICS OF CORRELATION

"The true Logic for this world is the Calculus of Probabilities, the only Mathematics for Practical Men."

— James Clerk Maxwell.

In an *Announcement* [1] issued April 29, 1916, by the *Office of Markets and Rural Organization* of the U. S. Department of Agriculture, there is a "*Review of Some of the Provisions of the Pending Cotton Futures Bill, H. R. 11861, and of Causes of Differences Between Prices of Middling Cotton in New York and Liverpool.*" Three valuable charts are given of fluctuations in different markets of prices of spot cotton and prices of cotton futures. One of the charts is described in these words: "Chart 3 shows the variation in the prices of futures [2] on the cotton exchanges at New York and New Orleans, as compared with the price of Middling as determined by averaging the quotations obtained from the designated spot markets, as follows: Norfolk, Augusta, Savannah, Montgomery, New Orleans, Memphis, Little Rock, Dallas, Houston, and Galveston. The chart

[1] *Service and Regulatory Announcements*, No. 9.

[2] The meaning of "futures" throughout the investigation is given in a description of the statistical data: "The future quotation for each day is always that for contracts which are to be fulfilled in the *current* month. During the last five days of a month, when contracts for the present month are no longer traded in, contracts for the following month are substituted, as they may be considered essentially the *current* month, for such contracts may be purchased or sold and immediately fulfilled or closed." *Ibid.*, p. 101.

covers the time between February 15, 1915, and January 22, 1916." From a study of this and the other two charts, the writer of the Review concludes "that since the cotton futures Act[1] went into operation future quotations have fairly reflected spot values in both New York and New Orleans, and also in a general way over the entire South, and that the law has thus accomplished and is accomplishing the end for which it was enacted." *Ibid.*, p. 104.

With the question as to whether the cotton futures Act is doing the work for which it was enacted, we are not at present concerned, but we are interested in the statement of fact that "since the cotton futures Act went into operation future quotations have fairly reflected spot values in both New York and New Orleans, and also in a general way over the entire South." The official words are that "future quotations have fairly reflected spot values." Just what is meant by *fairly?* How can one measure the degree of association between futures and spot values? Or, to put the question in another form, suppose one knew the probable spot values in the South, how could one forecast the price of futures on the cotton exchanges at New York and New Orleans? These are types of problems which the statistical methods we are about to describe enable us to solve. Let us propose a definite problem and connect the exposition of the statistical methods with the solution of the problem:

On Figure 1, the two graphs record for an interval of 42 days, from September 11 to October 30, 1915, the fluctuating prices of average spots in the South on

[1] The Act of 1914.

FIGURE 1. — The prices of spot cotton in the South, —, and the prices of cotton futures in New York, - - -.

the ten markets which were enumerated above, and the fluctuating prices of futures on the New York exchange. The general trend of each series of figures shows an ascent to about the middle of the record and then a descent. Suppose we were to make allowance for the general trend in the two graphs, what would be the degree of connection between the fluctuations of the futures from their general trend and the fluctuations of the spots from their general trend? To be more definite, suppose we represent the general trend of each series of figures by a progressive average of five daily quotations; that is to say, suppose we place on both series for each day a mark indicating the mean of the respective quotations for the five days of which the given day is the middle day. We should then obtain for each series a number of points that would indicate its general trend. If we take the fluctuations of each series from its own general trend, we shall have the data for the problem which we propose to solve, namely, to ascertain the degree of association between the fluctuations of futures and the fluctuations of spots.

Figure 2 shows the actual quotations for futures on the New York exchange and the general trend of the figures when the general trend is derived from a progressive average of five daily quotations. Figures 1 and 2 exhibit data for only 42 days.[1]

[1] The data used by the Government office, covering records for 275 to 280 days, were kindly supplied to me by Mr. Charles J. Brand, *Chief* of the *Office of Markets and Rural Organization.* When we come to the application of our statistical methods we shall use all of the available data.

FIGURE 2. — The general trend of cotton futures in New York.

We pass now to the development of the mathematical theory of correlation.[1]

A Frequency Distribution

Statistical tables that show either the absolute or relative frequencies of observations for given types of measurements are called frequency tables, or frequency distributions. The accompanying Table 1 is a frequency distribution showing the absolute frequencies in the fluctuations of the average prices of spots from their general trend, the general trend being derived from a progressive five days average.

After the raw observations, for purposes of facility in the handling of the data, have been grouped into appropriate frequency distributions, the next step is to describe the distributions by the aid of the fewest possible measurements that will enable one to summarize the features of the distribution which, for the purpose in hand, are most important.

One of the most important summary descriptions of a frequency distribution is the mean value of the distribution. In the particular problem before us the mean value of the fluctuations of average spots from their general trend is the quantity that we wish to ascertain. This brings us to the first step in our mathematical work.

[1] I wish most gratefully to thank Professor Karl Pearson for the instruction that I received in his laboratory several years ago, and for the inspiration of his published works. To him, almost exclusively, I owe my knowledge of the theory of correlation. In beginning the study of Professor Pearson's writings, I received help from Professor G. U. Yule's article "On the Theory of Correlation," in the *Journal of the Royal Statistical Society*, December, 1897, and from Mr. W. Palin Elderton's treatise on *Frequency Curves and Correlation*.

TABLE 1. — Frequency Distribution of Fluctuations of the Prices of Average Spots from a Five Days Progressive Mean of Prices

Fluctuations of Average Spots (Cents)	Frequency (Number of Days on which the Fluctuations Occurred)
—.165 to —.135	3
—.135 to —.105	3
—.105 to —.075	4
—.075 to —.045	23
—.045 to —.015	55
—.015 to +.015	107
+.015 to +.045	54
+.045 to +.075	16
+.075 to +.105	7
+.105 to +.135	2
+.135 to +.165	1
Total	275

Theorem I. *The algebraic sum of the deviations of a series of magnitudes from their arithmetical mean value is zero.*

Let the magnitudes be $x_1, x_2, x_3, \ldots x_n$, N in number, and let their arithmetical mean value be $\bar{x}$. Then, by the definition of the arithmetical mean, we have

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots x_n}{N},$$

$$\therefore N\,\bar{x} = x_1 + x_2 + x_3 + \ldots x_n,$$

and $(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + \ldots (x_n - \bar{x}) = 0$. But the quantities on the left-hand side of the equation are the deviations of the magnitudes from the arithmetical mean of the magnitudes, and the sum of these deviations is proved to be zero. This theorem we shall use later on in our work.

Theorem II. *The arithmetical mean of a series of magnitudes is equal to any arbitrary quantity plus the mean of the deviations of the magnitudes from the arbitrary quantity.*

As before, let the magnitudes be $x_1, x_2, x_3, \ldots x_n$, and let P be the arbitrary quantity.

Then $\bar{x} = \frac{\Sigma(x)}{N}$, where $\Sigma(x)$ is put for the sum of the x's. Also we have

$x_1 = P + x_1'$, where x_1' is the deviation of x_1 from P;
$x_2 = P + x_2'$, where x_2' " " " " x_2 from P;
$x_3 = P + x_3'$, where x_3' " " " " x_3 from P;
... ..." " " " ... ...
... ..." " " " ... ...
$x_n = P + x_n'$, where x_n' " " " " x_n from P.

Therefore $\Sigma(x) = N\,P + \Sigma(x')$, and $\frac{\Sigma(x)}{N} = P + \frac{\Sigma(x')}{N}$ which is the proposition we had to prove.

We shall now apply this latter theorem to find the mean value of the fluctuations of average spots from their general trend. The data are given in Table 2.

Here, the arbitrary quantity from which the fluctuations are measured is zero. Column II gives the fluctuations measured from zero expressed in terms of the unit of grouping. According to Theorem II, the arithmetical mean is equal to the arbitrary quantity plus the mean of the deviations from the arbitrary quantity. Consequently, in this particular case, the arithmetical mean of the price fluctuations is (— .07) in units of grouping, or (— .002) in absolute units.

The Standard Deviation as a Measure of Dispersion

The arithmetical mean of the frequency distribution gives us one of the most important summary descriptions of the distribution: it gives the centre of density of the distribution. But in economic, as well as in most other, measurements it is extremely important to know how the several observations are grouped about the arithmetical mean of the measurements, and a coefficient showing the manner of grouping is a measure of dispersion. Just as we found that the arithmetical mean of the measurements gives us an idea of the centre of the density of the measurements, so, as a measure of dispersion, we might take the arithmetical mean of the deviations of the magnitudes from the mean of the observations. But if we followed this

TABLE 2. — Computation of the Mean of the Fluctuations of Average Spots from Their General Trend

I Fluctuations of Average Spots	II Fluctuations Expressed in Units of Grouping Unit = .03 x'	III Frequency f	IV Product of Column II by Column III fx'
—.15	—5	3	— 15
—.12	—4	3	— 12
—.09	—3	4	— 12
—.06	—2	23	— 46
—.03	—1	55	— 55
0		107	
+.03	+1	54	+ 54
+.06	+2	16	+ 32
+.09	+3	7	+ 21
+.12	+4	2	+ 8
+.15	+5	1	+ 5
	Totals	275	—140 +120 — — 20

The mean fluctuation from the general trend is, therefore, $\frac{-20}{275} = -.07$ in units of grouping, or (—.07) (.03) = —.002 in absolute units.

plan, we should meet with an embarrassing difficulty: The deviations of the measurements from the arithmetical mean are some of them positive and some of them negative, and if we take account of the signs of

the deviations, then, according to Theorem I, the sum of the deviations is zero. We therefore choose, as our measure of dispersion, the square-root of the mean square of the deviations about the arithmetical mean of the observations, and we call this measure of dispersion the standard deviation.

If we let σ represent the standard deviation, then, if

$$\begin{aligned} x_1 - \bar{x} &= X_1, \\ x_2 - \bar{x} &= X_2, \\ x_3 - \bar{x} &= X_3, \\ \cdot\cdot \quad \cdot\cdot\cdot &\quad \cdot\cdot\cdot \\ x_n - \bar{x} &= X_n, \end{aligned}$$

we shall have as the symbolic expression of the standard deviation

$$\sigma_x = \sqrt{\left\{\frac{\Sigma(X^2)}{N}\right\}}$$

Theorem III. *The square of the standard deviation of a series of magnitudes is equal to the mean square of the deviations of the magnitudes about an arbitrary quantity, minus the square of the difference between the arbitrary quantity and the mean of the magnitudes.*

As before, let the quantities be $x_1, x_2, x_3, \ldots x_n$ and their mean value be $\bar{x}$. Let the arbitrary quantity be P and let the difference between the arbitrary quantity and the mean be d_x, so that $\bar{x} = P + d_x$. Let the deviations of the quantities from the arithmetical mean be $X_1, X_2, X_3, \ldots X_n$ and their deviations from P be $x'_1, x'_2, x'_3, \ldots x'_n$. We shall then have

$$(1)\qquad \begin{cases} X_1 = x_1 - \bar{x}, \\ X_2 = x_2 - \bar{x}, \\ X_3 = x_3 - \bar{x}, \\ \cdots \quad \cdots \quad \cdots \\ X_n = x_n - \bar{x}; \end{cases}$$

$$(2)\qquad P = \bar{x} - d_x;$$

$$(3)\qquad \begin{cases} x_1' = x_1 - P = x_1 - (\bar{x} - d_x) = (x_1 - \bar{x}) + d_x = X_1 + d_x, \\ x_2' = x_2 - P = x_2 - (\bar{x} - d_x) = (x_2 - \bar{x}) + d_x = X_2 + d_x, \\ x_3' = x_3 - P = x_3 - (\bar{x} - d_x) = (x_3 - \bar{x}) + d_x = X_3 + d_x, \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ x_n' = x_n - P = x_n - (\bar{x} - d_x) = (x_n - \bar{x}) + d_x = X_n + d_x; \end{cases}$$

$$(4)\qquad \begin{cases} (x_1')^2 = (X_1 + d_x)^2 = X_1^2 + 2\, d_x\, X_1 + d_x^2, \\ (x_2')^2 = (X_2 + d_x)^2 = X_2^2 + 2\, d_x\, X_2 + d_x^2, \\ (x_3')^2 = (X_3 + d_x)^2 = X_3^2 + 2\, d_x\, X_3 + d_x^2, \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ (x_n')^2 = (X_n + d_x)^2 = X_n^2 + 2\, d_x\, X_n + d_x^2; \end{cases}$$

(5) Therefore $\Sigma(x')^2 = \Sigma(X^2) + 2d_x\Sigma(X) + Nd_x^2$.

But according to Theorem I, $\Sigma(X) =$ zero, and, consequently $\dfrac{\Sigma(X^2)}{N} = \dfrac{\Sigma(x')^2}{N} - d_x^2$. Since σ_x^2 is by definition equal to $\dfrac{\Sigma(X^2)}{N}$, we have $\sigma_x^2 = \dfrac{\Sigma(x')^2}{N} - d_x^2$, which was to be proved.

Corollary. *The mean square deviation about the arithmetical mean of the observations is less than the mean square deviation about any arbitrary quantity.*

We have just proved that $\dfrac{\Sigma(X^2)}{N} = \dfrac{\Sigma(x')^2}{N} - d_x^2$. The left-hand side of the equation is a positive quantity because it is a mean square. The right-hand side must,

therefore, also be a positive quantity, but it consists of the difference between two positive quantities, the greater of which is the mean square deviation about an arbitrary quantity. The same equation would hold no matter what the arbitrary quantity might be. Therefore the mean square deviation about the arithmetical mean is less than the mean square deviation about any arbitrary quantity.

We shall now use this theorem to calculate the value of σ_x for the fluctuations of the average prices of spot cotton from the general trend of prices. The data are given in Table 3.

Our mathematical theory of correlation is developed as an instrument to forecast economic events. We may stop for a moment, therefore, to consider the bearing of our results thus far upon the problem of forecasting.

Figure 3 shows a smooth curve [1] passing closely to the broken line representing the frequency distribution of the fluctuations of average spot prices from their general trend. This curve is a symmetrical curve in the sense that the two sides of the figure are similarly disposed with reference to the maximum ordinate. If, for instance, the right-hand side of the figure were made to revolve about the maximum ordinate and be placed upon the left-hand side, the two parts of the curve would be congruent. This symmetrical curve is called the normal, or, sometimes, the Gaussian curve, after the author of *Theoria Motus Corporum Coelestium*, who was one of the first to investigate its properties. If we represent by x the deviations of the abscissas from

[1] In fitting the smooth curve to the data, the value of σ was computed with Sheppard's correction.

TABLE 3. — Computation of the Standard Deviation of the Fluctuations of Average Spots from the General Trend

I Fluctuations of Average Spots	II Fluctuations Expressed in Units of Grouping Unit = .03 x'	III Frequency f	IV Product of Column II by Column III fx'	V $(x')^2$	VI $f(x')^2$
—.15	— 5	3	— 15	25	75
—.12	— 4	3	— 12	16	48
—.09	— 3	4	— 12	9	36
—.06	— 2	23	— 46	4	92
—.03	— 1	55	— 55	1	55
0		107			
+.03	+ 1	54	+ 54	1	54
+.06	+ 2	16	+ 32	4	64
+.09	+ 3	7	+ 21	9	63
+.12	+ 4	2	+ 8	16	32
+.15	+ 5	1	+ 5	25	25
	Totals	275	— 140 + 120 — 20		544

According to the symbols in the text d_x is the mean deviation from the arbitrary origin, and, consequently, in this particular case, $d_x = \frac{-20}{275} = -.0727$, and $d_x^2 = .005285$. The mean square deviation about the arbitrary origin is $\frac{\Sigma f(x')^2}{N} = \frac{544}{275} = 1.978182$. By Theorem III, $\sigma_x^2 = \frac{\Sigma(X)^2}{N} = \frac{\Sigma f(x')^2}{N} - d_x^2$, and, consequently, $\sigma_x^2 = 1.978182 - .005285 = 1.972897$. Therefore $\sigma_x = \sqrt{1.972897} = 1.405$ in units of grouping,[1] or (1.405) (.03) = .042 in absolute units.

[1] The value of σ_x is here derived from the raw data, but in frequency curves of high contact, that is to say, in curves which tail off in the manner of Figure 3, the value of σ_x^2 is $\frac{\Sigma(X)^2}{N} - \frac{1}{12}$. This correction, which is known as Sheppard's correction, I have deliberately omitted because I wish to present the theory of correlation only in its bold outlines.

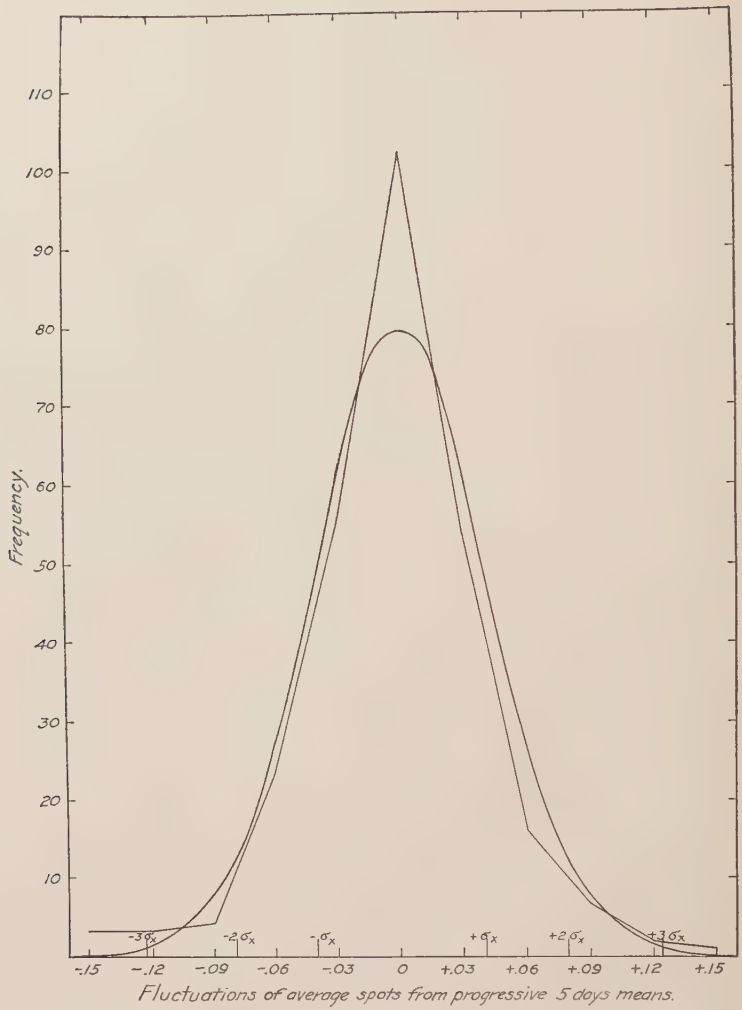

FIGURE 3. — The frequency distribution of the fluctuations of average spot prices from their general trend.

Equation to the smooth curve, $y = 79.81e^{-\frac{x^2}{.0034}}$.

the mean value of the distribution, and by y the ordinate corresponding to x, then the equation to the normal curve is $y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$, where N is the number of the observations, σ is the standard deviation of the observations, π is the ratio of the circumference of a circle to its diameter and is equal to 3.1416, and e is the base of Naperian logarithms and has the value 2.7183. If the distribution of the data is normal, all that is required to get the equation to the smooth Gaussian curve fitting the data is to substitute for N and σ, in the above equation, the values obtained for these constants from the concrete data of the problem.

From an investigation of the properties of the normal curve, it is found that in a perfectly normal distribution of data, 68 per cent of all the observations fall within $\pm \sigma$; that is to say, 68 per cent of the observations deviate less than plus σ or minus σ from the mean value of the frequency distribution. Furthermore, 95 per cent of all the observations fall within $\pm 2\sigma$, and 99.7 per cent of all the observations fall within $\pm 3\sigma$. On Figure 3, the distances of σ, 2σ, 3σ from the mean value of the distribution are indicated by special marks.[1]

We may now see why it was desirable to describe a

[1] As the normal curve is used so constantly in mathematical computations, tables have been constructed showing the proportion of cases included between the mean and a deviation from the mean of any multiple or submultiple of the standard deviation. One of the best compilations is Sheppard's *Tables of the Probability Integral*, which is contained in Professor Pearson's *Tables for Statisticians and Biometricians*. By the aid of this Table, after we have computed the mean and standard deviation of a given normal distribution, we can obtain the probability of any deviation from the mean value.

frequency distribution by its mean and its standard deviation. In any system of forecasting economic events, it is clearly of first importance to predict the events with the greatest possible precision, which is equivalent to reducing as far as possible the scatter or standard deviation of the predicted events. By taking the arithmetical mean about which to measure the scatter, we have in the standard deviation, according to the Corollary of Theorem III, a measure of dispersion which is less than the root-mean-square deviation about any other arbitrary quantity. Moreover, by the help of the *Tables of the Probability Integral*, when we use σ as the measure of scatter, we have at once, in case our frequency distributions are approximately normal, a numerical measure of the precision of our forecasts. This point will be illustrated further on in our work.

The Fitting of Straight Lines to Data

We have thus far described the mean and the standard deviation of a simple frequency distribution, and the particular distribution which we used as an illustration was the fluctuation of the average prices of spot cotton in the South from the general trend of spot prices, the general trend being derived from a five days progressive mean. What we did for the spot prices we could do for the prices of futures on the New York exchange. We should then be brought a step nearer to our concrete problem, which is to measure the degree of correlation between the fluctuations of New York futures and the fluctuations of average spots in the South.

Figure 4 displays the relation between the fluctuations of New York futures and the fluctuations of average spots in the South. A diagram like Figure 4 showing the relation between the variations of two variables is called a scatter diagram. From the general sweep of the scatter diagram, it is clear that, as the fluctuations of average spots increase or decrease, the fluctuations of New York futures increase or decrease. There is obvious association between the two variables, and the substance of our problem is to measure the degree of the correlation and to find the statistical law descriptive of the manner in which the one variable is connected with the other.

On the scatter diagram of Figure 4 the observations are represented by points.[1] Figure 5 describes the data of Figure 4 in a way to exhibit more clearly the nature of the correlation of the two variables. For each typical value of x in Figure 4 there is a corresponding array of y's, and Figure 5 shows, for each type-value of x, the mean value of the corresponding array of y's. It is clear that if we could fit properly a straight line to the points of Figure 5, the equation to the straight line would give us the statistical law connecting the changes of the two variables.

To cover the steps in the development of our method let us first recall that the trigometric tangent of an acute angle in a right-angle triangle is equal to the ratio of the side opposite the acute angle to the side

[1] There are 275 observations, but as a great many of them have the same values some of the representative points are superposed upon others. In making the computations that follow in the text each point is regarded as having an importance proportionate to the number of observations that it represents.

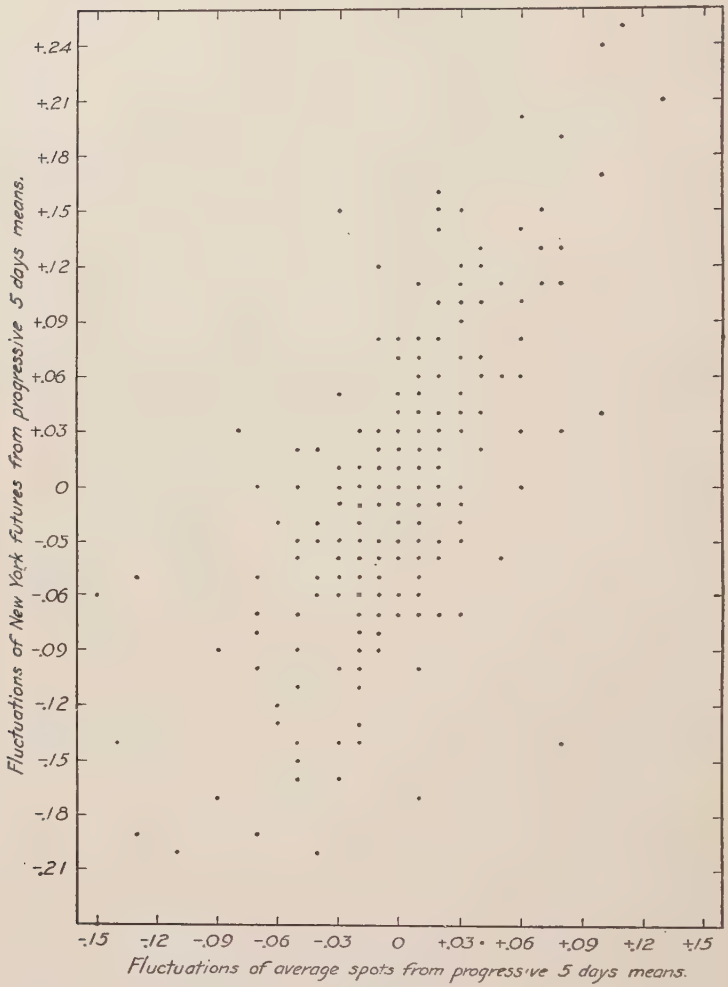

FIGURE 4. — Scatter diagram showing the relation between the fluctuations of futures and of spots.

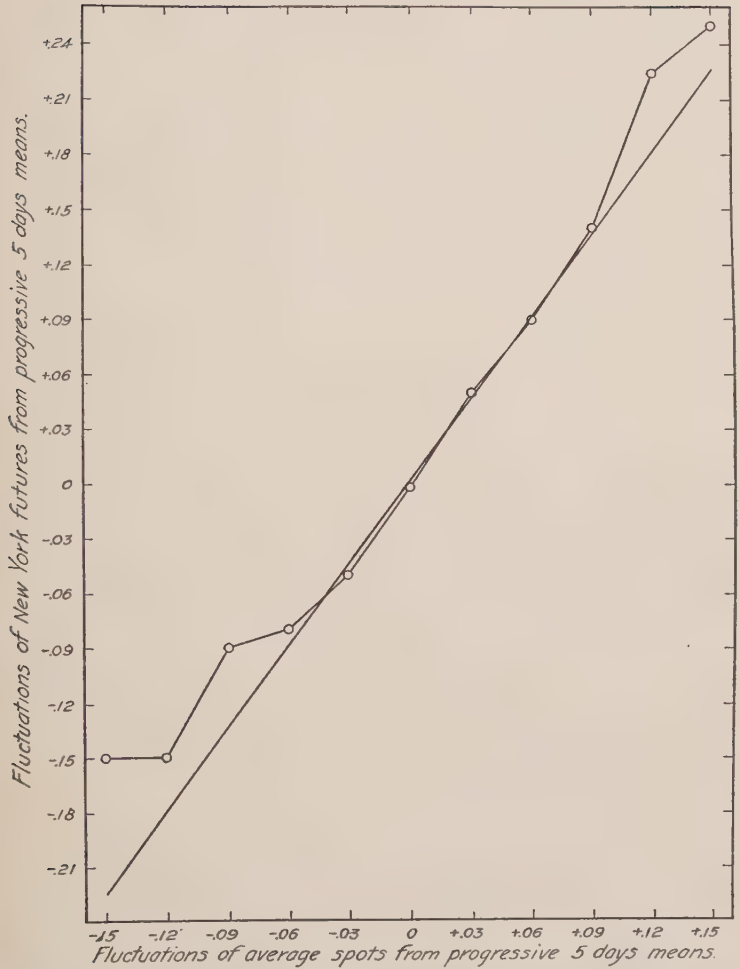

FIGURE 5. — The statistical law connecting the fluctuations of futures and of spots.

Equation to the straight line, $y = 1.50x + .002$, origin at (0, 0).

adjacent to the acute angle. For example, in the right-angle triangle, Figure 6, $\tan \alpha = \frac{BC}{AC}$. Let us further recall that the equation to a straight line may be put into the form $y = mx + b$, where m is the tangent of the angle which the line makes with the axis of x, and b is the intercept on the axis of y. For example, in Figure 7, DE is the line whose equation is sought. Let B represent a point on the line with coördinates (x, y). Then $y = BC + CF = AC \tan \alpha + b = x \tan \alpha + b = mx + b$. In the straight line corresponding to this equation, $y = mx + b$, the slope of the line will vary with the sign and magnitude of m, and the position of the line with reference to the axes of coördinates will vary with the sign and magnitude of b.

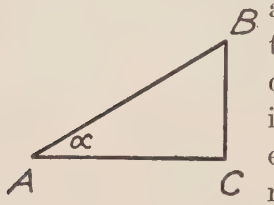

FIGURE 6.

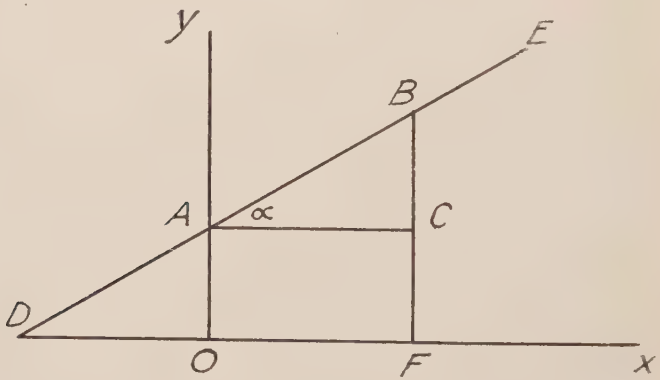

FIGURE 7.

The statistical problem is to find the values of m and b from the concrete data of the scatter diagram so that

the resulting straight line will give the best fit to the data. The expression "best fit" is seldom defined. Its significance varies with the problem in hand and it generally means a fit which is convenient and which, for the problem to be solved, gives satisfactory results.[1] The principle upon which the values of m and b are determined is so to choose m and b as to make the mean square deviation of the observations from the resulting straight line a minimum. The pertinency of this principle for our problem of forecasting is plain, because we have already learned that when observations are distributed according to the normal law, the *Tables of the Probability Integral* enable us to compute the probability of a deviation equal to any multiple or submultiple of the root-mean-square deviation. Moreover, as in all problems of forecasting it is desirable to have the root-mean-square deviation as small as possible, it is obvious that a straight line which fits given data so as to make the mean square deviation of the points from the straight line a minimum is, for the problem in hand of forecasting one variable from a knowledge of the other, a good fit to the data.

In Figure 5 let the abscissas of the series of points be x_1, x_2, x_3, . . . , and let the corresponding ordinates be $\bar{y}_{x_1}$, $\bar{y}_{x_2}$, $\bar{y}_{x_3}$, . . . Each of the ordinates, we recall, is the mean of the array of points in Figure 4 corresponding to the typical value of x. Suppose that $\bar{y}_{x_1}$ is determined from n_{x_1} points; $\bar{y}_{x_2}$ from n_{x_2} points; $\bar{y}_{x_3}$ from n_{x_3} points; and so on, for the other values of $\bar{y}_x$. Then $n_{x_1} + n_{x_2} + n_{x_3} + \ldots = N$, which is the total number of observa-

[1] See "The Statistical Complement of Pure Economics," *Quarterly Journal of Economics*, November, 1908, pp. 18 to 23.

tions. Now the condition that the mean square deviation of the points in Figure 5 from the straight line shall be a minimum is, symbolically, that

$$\frac{\Sigma\{n_x(y - \bar{y}_x)^2\}}{N} \text{ shall be a minimum.}$$

Here y is the ordinate of the straight line; $\bar{y}_x$ is the ordinate of a point in Figure 5 corresponding to abscissa x; n_x is the number of observations or points in the given array; N represents the total number of observations; and Σ indicates the operation of summing all the terms contained within the parentheses.

To facilitate the following development, let us put

$$(1) \qquad \frac{V}{N} = \frac{\Sigma\{n_x(y - \bar{y}_x)^2\}}{N}$$

Since $y = mx + b$, substitute this value of y in (1). Then

$$(2) \qquad \frac{V}{N} = \frac{\Sigma\{n_x(mx + b - \bar{y}_x)^2\}}{N}$$

$$= \frac{\Sigma\{n_x(m^2x^2 + 2mbx + b^2 - 2mx\bar{y}_x - 2b\bar{y}_x + \bar{y}_x^2)\}}{N}$$

$$(3) \quad \frac{V}{N} = m^2\frac{\Sigma(n_x x^2)}{N} + 2mb\frac{\Sigma(n_x x)}{N} + b^2\frac{\Sigma(n_x)}{N} - 2m\frac{\Sigma(n_x x\bar{y}_x)}{N} - 2b\frac{\Sigma(n_x\bar{y}_x)}{N} + \frac{\Sigma(n_x\bar{y}_x^2)}{N}$$

Now $\frac{\Sigma(n_x x^2)}{N} = \sigma_x^2 + \bar{x}^2$, where $\bar{x}$ = the mean of the x's, and σ_x, the standard deviation of the x's. This follows from Theorem III; $\frac{\Sigma(n_x x)}{N} = \bar{x}$, by the definition of the arithmetical mean; $\frac{\Sigma(n_x)}{N} = \frac{N}{N} = 1$;

$\frac{\Sigma(n_x x\bar{y}_x)}{N}$ = the mean of the xy products. This may be seen to be true from the following:

$$\bar{y}_{x_1} = \frac{{}_1y_{x_1} + {}_2y_{x_1} + {}_3y_{x_1} + \ldots + {}_{n_{x_1}}y_{x_1}}{n_{x_1}}.$$

Consequently,

$$n_{x_1}(x_1\bar{y}_{x_1}) = x_{1\,1}y_{x_1} + x_{1\,2}y_{x_1} + x_{1\,3}y_{x_1} + \ldots + x_{1\,n_{x_1}}y_{x_1}.$$

But what is true of this particular array is true of all the arrays and since there are N products xy, the value of $\frac{\Sigma(n_x x\bar{y}_x)}{N}$ is the mean of the xy products. Let us call the mean of the xy products p_{xy}. Continuing the investigation of equation (3), we have, $\frac{\Sigma(n_x\bar{y}_x)}{N}$ = the mean of the several values of $\bar{y}_x = \bar{y}$, where $\bar{y}$ is the mean of all the ordinates.

$\frac{\Sigma(n_x\bar{y}_x^2)}{N}$ = (the standard deviation of the $\bar{y}_x$'s)$^2 + \bar{y}^2$. This follows from Theorem III. Let us put $\sigma_{\bar{y}_x}$ for the standard deviation of the $\bar{y}_x$'s.

If now we make the proper substitutions, we may write equation (3) as follows:

$$(4) \qquad \frac{V}{N} = m^2(\sigma_x^2 + \bar{x}^2) + 2mb\bar{x} + b^2 - 2mp_{xy} - 2b\bar{y} + \sigma_{\bar{y}_x}^2 + \bar{y}^2$$

Our problem is to find the values of m and b that will make $\frac{V}{N}$ a minimum.

Let e_1, e_2 be two quantities so small that, for the pur-

pose in hand, we may neglect their squares and products and write

$$(m + e_1)^2 = m^2 + 2me_1;\ (b + e_2)^2 = b^2 + 2be_2;$$
$$(m + e_1)\ (b + e_2) = mb + e_2m + e_1b.$$

Suppose that, when in equation (4) we put $(m + e_1)$ for m and $(b + e_2)$ for b, we get

$$(5)\quad \frac{V'}{N} = (m + e_1)^2\ (\sigma_x^2 + \bar{x}^2) + 2(m + e_1)\ (b + e_2)\bar{x}$$
$$+ (b + e_2)^2 - 2(m + e_1)\ p_{xy} - 2(b + e_2)\ \bar{y} + \sigma_{\bar{y}_x}^2 + \bar{y}^2;$$
$$= (m^2 + 2me_1)\ (\sigma_x^2 + \bar{x}^2) + 2(mb + e_2m + e_1b)\ \bar{x} +$$
$$(b^2 + 2be_2) - 2(m + e_1)\ p_{xy} - 2\ (b + e_2)\ \bar{y} + \sigma_{\bar{y}_x}^2 + \bar{y}^2.$$

Let us consider a little more concretely the meaning of equations (4) and (5). Equation (4) indicates that if, in the straight line $y = mx + b$, we assign any given values to m and b, then the mean square of the deviations of the points from the straight line is equal to $\frac{V}{N}$. Equation (5) indicates that if we change the value of m in (4) to $(m + e_1)$ and the value of b in (4) to $(b + e_2)$ where e_1 and e_2 are very small quantities, then the mean square of the deviations of the points from the straight line $y = (m + e_1)\ x + (b + e_2)$ is given by $\frac{V'}{N}$. If m and b in (4) are so determined that $\frac{V}{N}$ is a minimum, then $\frac{V'}{N}$ is greater than $\frac{V}{N}$, and, by subtracting (4) from (5), we have

$$(6)\quad \frac{V' - V}{N} = 2me_1\ (\sigma_x^2 + \bar{x}^2) + 2(e_2m + e_1b)\ \bar{x} + 2be_2$$
$$- 2e_1p_{xy} - 2e_2\bar{y};$$
$$= 2e_1\ \{m(\sigma_x^2 + \bar{x}^2) + b\bar{x} - p_{xy}\} + 2e_2\ \{m\bar{x} + b - \bar{y}\}.$$

But when m and b are so determined as to render $\frac{V}{N}$ a minimum, and e_1 and e_2 are very small, $\frac{V'}{N}$ and $\frac{V}{N}$ are, for practical purposes, equal and equation (6) may be put into the form

$$(7)\quad 2e_1\{m(\sigma_x^2+\bar{x}^2)+b\bar{x}-p_{xy}\}+2e_2\{m\bar{x}+b-\bar{y}\}=0.$$

In order for equation (7) to be a true equation, a sufficient condition is that the coefficients of $2e_1$ and $2e_2$ shall each be zero; that is

$$\text{(i)}\quad m(\sigma_x^2+\bar{x}^2)+b\bar{x}-p_{xy}=0;$$
$$\text{(ii)}\quad m\bar{x}+b-\bar{y}=0.$$

Solve these two equations for m and b. Multiply (ii) by $\bar{x}$ and subtract the result from (i). We get $m\sigma_x^2-p_{xy}+\bar{x}\bar{y}=0$, and, consequently, $m=\frac{p_{xy}-\bar{x}\bar{y}}{\sigma_x^2}$. Substitute this value of m in (ii) and solve the resulting equation for b. We obtain $b=\bar{y}-\frac{p_{xy}-\bar{x}\bar{y}}{\sigma_x^2}\bar{x}$. If we substitute these values of m and b in the equation to the straight line, $y=mx+b$, we get

$$(8)\qquad y=\frac{p_{xy}-\bar{x}\bar{y}}{\sigma_x^2}x+\left\{\bar{y}-\frac{p_{xy}-\bar{x}\bar{y}}{\sigma_x^2}\bar{x}\right\}.$$

This is the equation to the straight line that makes $\frac{V}{N}$, the mean square deviation of the points from the line, a minimum; it is the straight line that fits best the data.

If our sole purpose were to find the line fitting best any given data, we might stop here. We should then

compute from the given data the values of the constants in equation (8), and, by substituting these values in that equation, obtain the equation in its numerical form. Our problem, however, is not completely solved by finding the equation connecting the two variables x and y. We wish to know, in any given case, how closely the two variables are associated. In the particular case which we have taken to illustrate our mathematical methods, we wish not only to know the equation connecting the fluctuations of New York futures from their general trend with the fluctuations of the average prices of spot cotton from their general trend, but we wish to know how closely the prices of futures and the prices of spot cotton are connected. To approach this last problem we simplify equation (8).

We have agreed to call $\bar{x}$ the mean of the x's in the scatter diagram, and $\bar{y}$, the mean of the y's. Suppose we call the point in the scatter diagram whose coördinates are $(\bar{x}, \bar{y})$ the mean of the system of points, and inquire whether the straight line described by equation (8) passes through the mean of the system of points. If the line passes through this point, the coördinates of the point $(\bar{x}, \bar{y})$ must satisfy the equation. Substitute $\bar{x}$, $\bar{y}$ respectively for x and y in (8). We obtain

$$(9) \qquad \bar{y} = \frac{p_{xy} - \bar{x}\bar{y}}{\sigma_x^2}\,\bar{x} + \left\{\bar{y} - \frac{p_{xy} - \bar{x}\bar{y}}{\sigma_x^2}\,\bar{x}\right\}, \text{ or,}$$

$$\bar{y} - \frac{p_{xy} - \bar{x}\bar{y}}{\sigma_x^2}\,\bar{x} = \bar{y} - \frac{p_{xy} - \bar{x}\bar{y}}{\sigma_x^2}\,\bar{x}.$$

But this is a true and identical equation, and consequently the line described by equation (8) passes through the mean of the system of points on the scatter

diagram, that is to say, the point whose coördinates are $(\bar{x}, \bar{y})$.

The fact that the best-fitting line passes through the point $(\bar{x}, \bar{y})$ enables us to simplify equation (8). By transposing we may write (8) as follows:

$$(10) \qquad (y - \bar{y}) = \frac{p_{xy} - \bar{x}\bar{y}}{\sigma_x^2}(x - \bar{x}).$$

The quantity $(y - \bar{y})$ is the deviation of the ordinate of the best-fitting straight line from the mean of the y's, and may be represented by Y; the quantity $(x - \bar{x})$ is the deviation of the abscissa of the line from the mean of the x's, and may be represented by X. Since, as we have just proved, the line passes through the point $(\bar{x}, \bar{y})$, if we transfer the origin from zero to the point $(\bar{x}, \bar{y})$, equation (10) may be written

$$(11) \qquad Y = \frac{p_{xy} - \bar{x}\bar{y}}{\sigma_x^2} X.$$

The effect of transferring the origin to the point $(\bar{x}, \bar{y})$ is to get rid of the value of b in the equation to the straight line.

We shall now examine the quantity $\frac{p_{xy} - \bar{x}\bar{y}}{\sigma_x^2} = m$, which appears in both (10) and (11). We know that p_{xy} is the mean value of the products xy. Let us define a new quantity π_{xy} to be the mean product of the deviations of x and y from their respective means. Then, by definition,

$$\pi_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{N} = \frac{\Sigma(xy)}{N} - \bar{x}\frac{\Sigma(y)}{N} - \bar{y}\frac{\Sigma(x)}{N} + \bar{x}\bar{y}.$$

But $$\frac{\Sigma(xy)}{N} = p_{xy}; \frac{\Sigma(y)}{N} = \bar{y}; \frac{\Sigma(x)}{N} = \bar{x}.$$

Therefore $\pi_{xy} = p_{xy} - \bar{x}\bar{y}$, and we may write $m = \frac{p_{xy} - \bar{x}\bar{y}}{\sigma_x^2} = \frac{\pi_{xy}}{\sigma_x^2}$. Make this substitution in equations (10) and (11) and we get

(12) $$(y - \bar{y}) = \frac{\pi_{xy}}{\sigma_x^2}(x - \bar{x});$$

(13) $$Y = \frac{\pi_{xy}}{\sigma_x^2} X.$$

If, as a further step, we define r to be a quantity such that $r = \frac{\pi_{xy}}{\sigma_x \sigma_y}$, then, by substituting in (12) and (13), we may write the equation to the best-fitting straight line in either of the following forms:

(14) $$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}).$$

(15) $$Y = r\frac{\sigma_y}{\sigma_x} X.$$

The quantity r in these equations is called the coefficient of correlation.

The Coefficient of Correlation

An inspection of equations (14) and (15) shows that in order to secure the best fit of a straight line to given data, all that is necessary is to compute from the data the values of $\bar{x}$, $\bar{y}$, σ_x, σ_y, r, and to make the proper substitutions in (14) and (15). We have already discussed methods of computing $\bar{x}$, $\bar{y}$, σ_x, σ_y, and we now

reach the question of the best method of computing r. As we have just seen, $r = \frac{\pi_{xy}}{\sigma_x \sigma_y} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{N\sigma_x \sigma_y}$, and if we were indifferent to the labor of computation, we might use this formula to ascertain, in any concrete case, the value of r. We, however, found methods for computing $\bar{x}$, $\bar{y}$, σ_x, σ_y by working with deviations from arbitrary quantities as origins, and we now proceed to develop a method of computing r by retaining the same arbitrary origins which we used in calculating the means and the standard deviations. We wish to find $\frac{\Sigma(x - \bar{x})(y - \bar{y})}{N}$ and to derive its value by working with deviations of the x's and y's from arbitrary origins.

Theorem IV. *The mean product of the deviations of two correlated variables from their respective arithmetical means is equal to the mean product of the deviations of the two variables from arbitrary origins, minus the difference between the arbitrary origin and the mean of the one variable multiplied by the difference between the arbitrary origin and the mean of the second variable.*

Let the observations be (x_1, y_1); (x_2, y_2); (x_3, y_3) . . . (x_n, y_n). Let P be the arbitrary origin from which we measure the deviations of the x's, and Q be the arbitrary origin from which we measure the deviations of the y's. Let the deviations of the x's from P be represented by x' and the deviation of the y's from Q be represented by y'. Let the deviations of the x's from $\bar{x}$ be represented by X, and the deviations of the y's from $\bar{y}$ be represented by Y. If we put $\bar{x} - P = d_x$, and $\bar{y} - Q = d_y$, our Theorem IV is that

$$\frac{\Sigma(x - \bar{x})(y - \bar{y})}{N} = \frac{\Sigma(x'y')}{N} - d_x d_y.$$

We have

(16)

$$x_1' = x_1 - P = x_1 - (\bar{x} - d_x) = (x_1 - \bar{x}) + d_x = X_1 + d_x,$$
$$x_2' = x_2 - P = x_2 - (\bar{x} - d_x) = (x_2 - \bar{x}) + d_x = X_2 + d_x,$$
$$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$$
$$x_n' = x_n - P = x_n - (\bar{x} - d_x) = (x_n - \bar{x}) + d_x = X_n + d_x.$$

Similarly,

$$y_1' = y_1 - Q = y_1 - (\bar{y} - d_y) = (y_1 - \bar{y}) + d_y = Y_1 + d_y,$$
$$y_2' = y_2 - Q = y_2 - (\bar{y} - d_y) = (y_2 - \bar{y}) + d_y = Y_2 + d_y,$$
$$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$$
$$y_n' = y_n - Q = y_n - (\bar{y} - d_y) = (y_n - \bar{y}) + d_y = Y_n + d_y.$$

Therefore,

$$x_1'y_1' = (X_1 + d_x)(Y_1 + d_y) = X_1Y_1 + d_yX_1 + d_xY_1 + d_xd_y,$$
$$x_2'y_2' = (X_2 + d_x)(Y_2 + d_y) = X_2Y_2 + d_yX_2 + d_xY_2 + d_xd_y,$$
$$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$$
$$x_n'y_n' = (X_n + d_x)(Y_n + d_y) = X_nY_n + d_yX_n + d_xY_n + d_xd_y.$$

Summing both sides of the equation, we get

$$\Sigma(x'y') = \Sigma(XY) + d_y\Sigma(X) + d_x\Sigma(Y) + Nd_xd_y.$$

But, according to Theorem I, $\Sigma(X) = \Sigma(Y) = 0$, and, consequently,

$$\Sigma(x'y') = \Sigma(XY) + Nd_xd_y, \text{ or,}$$

$$(17) \quad \frac{\Sigma(XY)}{N} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{N} = \frac{\Sigma(x'y')}{N} - d_xd_y.$$

This formula gives us a method of computing the factor $\frac{\Sigma(x - \bar{x})(y - \bar{y})}{N}$ in the value of r, the formula for which we know is, $r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{N\sigma_x\sigma_y}$.

TABLE 4. — Correlation Table Showing the Relation Between Fluctuations of the Prices of New York Cotton Futures and Fluctuations of the Prices of Spot Cotton on the Exchanges of the South

Fluctuations of New York futures from the progressive means of 5 days	Fluctuations of average spots from the progressive means of 5 days											
	—.165 to —.135	—.135 to —.105	—.105 to —.075	—.075 to —.045	—.045 to —.015	—.015 to +.015	+.015 to +.045	+.045 to +.075	+.075 to +.105	+.105 to +.135	+.135 to +.165	Totals
—.275 to —.225	25 1 (25)						— 5 1 (— 5)					2
—.225 to —.175		16 2 (32)		8 1 (8)	4 1 (4)							4
—.175 to —.125	15 1 (15)		9 2 (18)	6 4 (24)	3 6 (18)	0 1 (0)						14
—.125 to —.075			6 1 (6)	4 8 (32)	2 7 (14)	0 6 (0)						22
—.075 to —.025	5 1 (5)	4 1 (4)		2 6 (12)	1 25 (25)	0 29 (0)	— 1 5 (— 5)	— 2 1 (— 2)				68
—.025 to +.025				0 4 (0)	0 12 (0)	0 39 (0)	0 9 (0)					64
+.025 to +.075			— 3 1 (— 3)		— 1 4 (— 4)	0 25 (0)	1 19 (19)	2 4 (8)	3 2 (6)			55
+.075 to +.125						0 7 (0)	2 14 (28)	4 7 (28)	6 1 (6)			29
+.125 to +.175							3 6 (18)	6 3 (18)	9 2 (18)			11
+.175 to +.225								8 1 (8)	12 1 (12)	16 1 (16)		3
+.225 to +.275									15 1 (15)	20 1 (20)	25 1 (25)	3
Totals	3	3	4	23	55	107	54	16	7	2	1	275
Means	—.15	—.15	—.09	—.08	—.05	—.002	+.05	+.09	+.14	+.225	+.25	—.001

The data in Table 4 will serve to illustrate the method of computing the coefficient of correlation.[1] We have proved that $r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{N\sigma_x\sigma_y}$, and that $\frac{\Sigma(x-\bar{x})(y-\bar{y})}{N} = \frac{\Sigma(x'y')}{N} - d_x d_y$; and we recall that $d_x = \bar{x} - P$, where P is the arbitrary origin from which x' is measured, and $d_y = \bar{y} - Q$ where Q is the arbitrary origin from which y' is measured. In the correlation table which is given in Table 4, P is the origin from which are measured the fluctuations of the average spots about the progressive means of 5 days, and is taken at the point zero, which lies mid-way between (— .015) and (+ .015). The values of x', the fluctuations of average spots, are the distances to the right and to the left of the arbitrary origin, and the sign of x' is positive or negative according as the distance is to the right or to the left of the arbitrary origin. In a similar manner, the arbitrary origin Q, from which are measured the fluctuations of New York futures about the progressive means of 5 days, is taken at the point zero which lies mid-way between (— .025) and (+ .025). The fluctuations from Q, which are designated by y', are negative toward the upper end of the table and positive toward the lower end. Just as in the scatter diagram, which is given in Figure 4, the 275 observations were represented, according to their co-

[1] The method described in the text is the one most frequently required in actual experience. Where, however, the number of observations is small, which happens to be the case with a large part of the data in this Essay, a slight alteration of the procedure described in the text is necessary. A complete illustration of the method of correlation when the observations are few in number is given in Chapter III, Table 6.

ordinates, by points on the diagram, so in the correlation table each observation falls in some one of the cells composing the Table. The figure in the middle of the cell gives the number of observations in the cell; for example, in the upper left-hand corner there is one observation, which means that out of 275 days observation, there was one day when the fluctuation of average spots was between (— .165) and (— .135) from the general trend of spots, and the fluctuation of New York futures was between (— .275) and (— .225) from the general trend of New York futures. In the same cell in the upper left-hand corner of the correlation table there is above the figure 1 the figure 25, and below the figure 1, the figure (25). A similar arrangement is followed in all of the cells in which observations occur, and we now proceed to explain its meaning. The working unit in the classification of the fluctuations of spots is .03, and in the classification of the fluctuations of New York futures, it is .05. The range of the fluctuations of spots is from the mid-value of the first cell on the left to the mid-value of the last cell on the right, that is, from (— .15) to (+ .15), or, since the working unit of the x''s is .03, the range is from (— 5) to (+ 5) working units. Similarly, the range of the y''s is from (— .25) to (+ .25), or, since the working unit is .05, the range is from (— 5) to (+ 5) working units.

Returning now to the one observation in the upper left-hand corner of the correlation table, we find that its distance from the zero point of the x''s is (— 5) working units, and from the zero point of the y''s is also (— 5) working units. The product of these two, which is $x'y'$, is $(-5)(-5) = 25$, and this explains

the figure 25 at the top of this one cell. Since there is only one observation in this cell, if we weight the product 25 by 1 we get (25), which explains the figure (25) at the bottom of this particular cell. To summarize, the figure in the middle of the cell is the frequency of the observations; the figure at the top of the cell is the product $x'y'$ in working units; and the figure at the bottom of the cell is $x'y'$ weighted according to the number of observations in the cell. The heavy lines that pass from the top to the bottom, and from the left to the right of the correlation table divide the latter into four large divisions. All of the products in the cells of the upper left-hand and lower right-hand divisions are positive, and all of the products of the other two divisions are negative. If we sum all of the positive products separately and then all of the negative products, their difference will give us $\Sigma(x'y')$; and if we then divide this result by 275 we shall obtain $\frac{\Sigma(x'y')}{N}$. If we indicate by $\Sigma(+\,x'y')$ the sum of the positive products, and by $\Sigma(-\,x'y')$ the sum of the negative products, we find from Table 4 that $\Sigma(+\,x'y') = (25) + (15) + (5) + (32) + (4) + (18) + (6) + (8) + (24) + (32) + (12) + (4) + (18) + (14) + (25) + (19) + (28) + (18) + (8) + (28) + (18) + (8) + (6) + (6) + (18) + (12) + (15) + (16) + (20) + (25) = 487$; and $\Sigma(-\,x'y') = (-\,3) + (-\,4) + (-\,5) + (-\,5) + (-\,2) = -\,19$. Consequently, $\Sigma(x'y') = 487 - 19 = 468$, and $\frac{\Sigma(x'y')}{N} = \frac{468}{275} = 1.7018$. The quantity that we wish to determine next is $\frac{\Sigma(x - \bar{x})(y - \bar{y})}{N}$,

which we know is equal to $\frac{\Sigma(x'y')}{N} - d_x d_y$. We have found in the early part of this chapter that $d_x = -.073$ in working units; and just as we determined d_x we can, in a similar manner, determine d_y. The actual computation shows that $d_y = -.026$ in working units. Consequently, $d_x d_y = (-.073)(-.026) = .0019$, and $\frac{\Sigma(x'y')}{N} - d_x d_y = 1.7018 - .0019 = 1.6999$, which is, therefore, the value of $\frac{\Sigma(x - \bar{x})(y - \bar{y})}{N}$. But the coefficient of correlation r is equal to $\frac{\Sigma(x - \bar{x})(y - \bar{y})}{N\sigma_x\sigma_y}$, and since, in working units, $\sigma_x = 1.405$ and $\sigma_y = 1.694$, we have $r = \frac{1.6999}{2.3801} = .714$.

Only one other short step is needed to get the equation to the straight line that fits best the data. (See Figure 5.) We know that the equation to the best-fitting straight line is $(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$, and all that is necessary is to substitute for the symbols their numerical values: $\bar{x} = -.002$; $\bar{y} = -.001$, $\sigma_x = (1.405)(.03) = .042$; $\sigma_y = (1.694)(.05) = .085$; $r = .714$. The proper substitution gives for the equation to the line,[1] $y = 1.45x + .002$.

The equation $y = 1.45x + .002$ gives the law of the association of the price of New York futures with the price of average spots in the South. Whatever may be the value of x, which is the fluctuation of average spots

[1] The slight difference between this equation and the one given in Figure 5 is due to the fact that in computing the latter equation Sheppard's correction was used in getting the values of σ_x and σ_y.

from their general trend, the above equation enables one to compute the most probable value of y, which is the fluctuation of New York futures from their general trend. For example, if x should be equal to $(+ .15)$, the most probable value of y is, $y = 1.45(.15) + .002 = .22$.

The geometrical significance of r. We have proved that the equation to the best-fitting straight line is $Y = r\frac{\sigma_y}{\sigma_x}X$. Suppose we write this equation in the following form:

$$\left(\frac{Y}{\sigma_y}\right) = r\left(\frac{X}{\sigma_x}\right). \tag{18}$$

This expression enables us to form a picture of the geometrical significance of r. Equation (18) shows that if the X's are expressed in terms of σ_x and the Y's in terms of σ_y, then r is the tangent of the angle which the straight line makes with the axis of $\left(\frac{X}{\sigma_x}\right)$. The value of r shows the proportional change in $\left(\frac{Y}{\sigma_y}\right)$ corresponding to a unit change in $\left(\frac{X}{\sigma_x}\right)$.

The limits to the value of r are ± 1. The equation to the best-fitting straight line is $Y = r\frac{\sigma_y}{\sigma_x}X$. Let the N observations be $(X_1, Y_1); (X_2, Y_2); \ldots (X_n, Y_n)$. Then, by the definition of V, we have

$$V = (Y_1 - r\frac{\sigma_y}{\sigma_x}X_1)^2 + (Y_2 - r\frac{\sigma_y}{\sigma_x}X_2)^2 + \ldots + (Y_n - r\frac{\sigma_y}{\sigma_x}X_n)^2,$$

$$= \Sigma(Y^2) - 2r\frac{\sigma_y}{\sigma_x}\Sigma(XY) + r^2\frac{\sigma_y^2}{\sigma_x^2}\Sigma(X^2).$$

But $\Sigma(Y^2) = N\sigma_y^2$; $\Sigma(X^2) = N\sigma_x^2$; $\Sigma(XY) = Nr\sigma_x\sigma_y$.

Consequently,

$V = N\sigma_y^2 - 2Nr^2\sigma_y^2 + Nr^2\sigma_y^2 = N\sigma_y^2(1 - r^2)$, and therefore,

$$\frac{V}{N} = \sigma_y^2(1 - r^2). \tag{19}$$

But $\frac{V}{N}$, being the mean square deviation of the observations from the straight line, is a positive quantity. Therefore, r cannot exceed $(+1)$ nor be less than (-1). The fact which was brought out a moment ago, namely, that r is the tangent of the angle which one straight line makes with another, shows that the value of r may be positive or negative according to the inclination of the line.

The use of r in the problem of forecasting. Equation (19) gives us a formula which is of the very first importance in our effort to forecast economic events. We have $\frac{V}{N} = \sigma_y^2(1 - r^2)$, and this is the measure of the mean square deviation of the points from the straight line that fits best the observations. When $r = (+1)$ or (-1), $\frac{V}{N}$ equals zero; all of the points lie on the straight line; and, by means of the equation to the straight line, we can predict exactly the value of y corresponding to a given value of x. But it is very seldom that $r = \pm 1$, and when r lies between these two limiting values, we can still forecast results with a knowledge of the probabilities in favor of the forecast. Let us put $S = \sqrt{\left(\frac{V}{N}\right)} = \sigma_y\sqrt{1 - r^2}$. We know from the

Table of the Probability Integral that when the distribution of the points about the straight line is normal, 99.7 out of 100 observations lie within a deviation from the straight line equal to $\pm 3S$; 95 out of 100 lie between $\pm 2S$; and 68 out of 100 lie between $\pm S$. The equation to the best-fitting straight line enables us to compute the most probable value of y corresponding to a given value of x; the value of S enables us to say within what limits any proportion of the actual observations are scattered about the straight line. The coefficient of correlation r is the coefficient which we have been seeking as a measure of the degree of association between two variables. Where the association between the variables is perfect, $r = \pm 1$, $S = \sigma_y\sqrt{1 - r^2} = 0$, and from the knowledge of the one variable we can, by means of the equation to the best-fitting straight line, forecast the other variable with perfect accuracy. When the association between the two variables is not perfect, r falls between the limiting values ± 1, and $S = \sigma_y\sqrt{1 - r^2}$ shows the accuracy with which, using the equation to the best-fitting straight line, the magnitude of the one variable may be predicted from a knowledge of the other.

We may illustrate these points by the problem of the relation between New York futures and average spot values in the South. We have found that the best-fitting straight line connecting the fluctuations in New York futures with the fluctuations in the price of spot cotton is $y = 1.45x + .002$. For any given value of x, representing the fluctuation in the price of spot cotton, we can predict, by means of this formula, the most probable fluctuation in the price of New York futures.

We are, however, not content to forecast the most probable values of y, but we wish to know, in addition, the degree of accuracy of the forecasts. The formula that has just been developed supplies an answer to this latter question. Since $r = .714$ and $\sigma_y = .085$, therefore $S = \sigma_y\sqrt{1 - r^2} = .06$, and from what we have learned about the significance of S, we know that, when we use the formula $y = 1.45x + .002$ as a prediction formula, in 99.7 per cent of all the forecasts the error will be less than $\pm 3S$; in 95 per cent of all the forecasts, the error will be less than $\pm 2S$; and in 68 per cent of all the forecasts, the error will be less than $\pm S$.

In beginning this chapter we referred to the official statement that "since the cotton futures Act went into operation, future quotations have fairly reflected spot values in both New York and New Orleans, and also in a general way over the entire South." We made the comment: "Just what is meant by *fairly?* How can one measure the degree of association between futures and spot values? Or, to put the question in another form, suppose one knew the probable spot values in the South, how could one forecast the price of futures" on the cotton exchange at New York? All of these questions may now be answered in a definite, numerical way: the degree of association between futures in New York and spot values in the South is measured by $r = .714$; the formula by which futures may be predicted from the knowledge of spot values is $y = 1.45x + .002$; and the error of the forecasts by means of this formula is measured by $S = .06$.

CHAPTER III

THE GOVERNMENT CROP REPORTS

"The consequences of false reports concerning the condition and prospective yield of the cotton crop alone may be very damaging. If there were no adequate Government crop-reporting service, and by misleading reports speculators should depress the price a single cent per pound, growers would lose $60,000,000 or more; if prices were improperly increased, the manufacturers and allied interests would be affected to a proportionate degree."

—*Circular 17, Bureau of Statistics, U. S. Department of Agriculture.*

The character and the aim of the official crop-reporting service, the definition and the use of technical terms, and the actual procedure in crop-forecasting have been described in publications of the Department of Agriculture.[1] The official documents might, of course, be summarized, but it seems advisable to quote in full those statements that have a bearing upon the subject of the present and subsequent chapters of this Essay.

The Character and the Aim of the Crop-Reporting Service

The Department of Agriculture "is said to have been conceived in the far-sighted wisdom of Washington, who, as President, suggested the organization of a branch of the National Government to care for the interests of the farmers; and, in the practical activity of Franklin, who, as agent of Pennsylvania in England sent home silkworm eggs and mulberry cuttings to start silk growing. But the

[1] I wish to acknowledge with hearty thanks the courteous helpfulness of the officials of the Department of Agriculture in supplying me with statistical material. Mr. Charles J. Brand, Mr. Leon M. Esterbrook, Mr. George K. Holmes, and Mr. Nat C. Murray were particularly generous in their assistance.

conception did not materialize into form until 1839, when, on the recommendation of the Hon. Henry L. Ellsworth, Commissioner of Patents, an appropriation of $1,000 was made by Congress for the 'collection of agricultural statistics, investigation promoting agricultural and rural economy, and the procurement of cuttings and seeds for gratuitous distribution among farmers.'

"An agricultural section was established in the Patent Office, and the collection of seeds and the publication of agricultural statistics and scientific articles on agricultural topics were placed directly under control of the Commissioner of Patents, at that time an official of the Department of State; the work continued under succeeding Commissioners of Patents until 1849, when the Department of the Interior was established, and the Patent Office, with its agricultural section, became a part of it. From that time until 1862, when the section was made a separate Department under a Commissioner of Agriculture, the agricultural work was done by the chief of the section of agriculture in the Patent Office, under the direction of the Commissioner of Patents.

"From 1862 until 1889, when the Department was raised to the dignity of a cabinet office, the work was prosecuted under the Commissioner of Agriculture, independently of the Department of the Interior.

"Under President Cleveland's first administration the Department became, on February 11, 1889, one of the Executive Departments of the Government." [1]

"The first enactment authorizing the collection of agricultural statistics [2] by the Department of Agriculture was the act, passed May 15, 1862, establishing the Department, 'the general design and duties of which shall be to acquire and to diffuse among the people of the United States information on subjects connected with agriculture, in the most general and comprehensive sense of the word.'

[1] "The United States Department of Agriculture," *Crop Reporter*, January, 1901, pp. 1–2.

[2] The Government publication from which the following quotations are made bears the title: *Government Crop Reports: Their Value, Scope, and Preparation*, and forms Circular 17, of the Bureau of Statistics, of the U. S. Department of Agriculture. As the date of the Circular is September 30, 1908, some points of detail may not now be accurate. But as our investigation will cover the quarter of a century, 1890–1914, what was said in 1908 will give a general idea of the organization and activity of the Department during the period under investigation.

The Commissioner was required by this act to 'procure and preserve all information concerning agriculture which he can obtain by means of books, correspondence, and by practical and scientific experiments, accurate records of which experiments shall be kept in his office, by the collection of statistics, and by any other appropriate means within his power.'

"The first appropriation for collecting agricultural statistics by the Department was provided for by the act of February 25, 1863, which was made in bulk for the work of the Department, amounting in all to $90,000. The then Commissioner of Agriculture allotted a part of this amount for collecting agricultural statistics, and appointed a statistician for that purpose. For the fiscal year ended June 30, 1865, the first distinct and separate provision was made for collecting agricultural statistics for information and reports, and the amount of $20,000 was appropriated.

"From an allotment of a few thousand dollars each year at first the crop-reporting service has been evolved, perfected, and enlarged into the Bureau of Statistics of this Department.

"The appropriation act for the Department of Agriculture for the fiscal year ended June 30, 1908, carried appropriations of about $220,000 for the Bureau of Statistics, and for the current year the appropriation has been increased to about $222,000. As the appropriations for the statistical and crop-reporting service have been gradually increased during the past several years, the field service and organization of the Bureau have been correspondingly enlarged.

"The Bureau of Statistics issues each month detailed reports relating to agricultural conditions throughout the United States, the data upon which they are based being obtained through a special field service, a corps of State statistical agents, and a large body of voluntary correspondents composed of the following classes: County correspondents, township correspondents, individual farmers, and special cotton correspondents.

"The special field service consists of seventeen traveling agents, each assigned to report for a separate group of States. These agents are especially qualified by statistical training and practical knowledge of crops. They systematically travel over the district assigned to them, carefully note the development of each crop, keep in touch with best informed opinion, and render written and telegraphic reports monthly and at such other times as required.

"There are forty-five State statistical agents, each located in a different State. Each reports for his State as a whole, and maintains a corps of correspondents entirely independent of those reporting directly to the Department at Washington. These State statistical correspondents report each month directly to the State agent on schedules furnished him. The reports are then tabulated and weighted according to the relative product or area of the given crop in each county represented, and are summarized by the State agent, who coördinates and analyses them in the light of his personal knowledge of conditions, and from them prepares his reports to the Department.

"There are approximately 2,800 counties of agricultural importance in the United States. In each the Department has a principal county correspondent who maintains an organization of several assistants. These county correspondents are selected with special reference to their qualifications and constitute an efficient branch of the crop-reporting service. They make the county the geographical unit of their reports, and, after obtaining data each month from their assistants and supplementing these with information obtained from their own observation and knowledge, report directly to the Department at Washington.

"In the townships and voting precincts of the United States in which farming operations are extensively carried on the Department has township correspondents who make the township or precinct the geographical basis of reports, which they also send directly to the Department each month.

"Finally, at the end of the growing season a large number of individual farmers and planters report on the results of their own individual farming operations during the year; valuable data are also secured from 30,000 mills and elevators.

"With regard to cotton, all the information from the foregoing sources is supplemented by that furnished by special cotton correspondents, embracing a large number of persons intimately concerned in the cotton industry; and, in addition, inquiries in relation to acreage and yield per acre of cotton are addressed to the Bureau of the Census's list of cotton ginners through the courtesy of that Bureau.

"Eleven monthly reports on the principal crops are received yearly from each of the special field agents, county correspondents, State statistical agents, and township correspondents, and one re-

port relating to the acreage and production of general crops annually from individual farmers.

"Six special cotton reports are received during the growing season from the special field agents, from the county correspondents, from the State statistical agents, and from township correspondents, and the first and last of these reports are supplemented by returns from individual farmers, special correspondents, and cotton ginners.

"In order to prevent any possible access to reports which relate to speculative crops, and to render it absolutely impossible for premature information to be derived from them, all of the reports from the State statistical agents, as well as those of the special field agents, are sent to the Secretary of Agriculture in specially prepared envelopes addressed in red ink with the letter 'A' plainly marked on them. By an arrangement with the postal authorities these envelopes are delivered to the Secretary of Agriculture in sealed mail pouches. These pouches are opened only by the Secretary or Assistant Secretary, and the reports, with seals unbroken, are immediately placed in the safe in the Secretary's office, where they remain sealed until the morning of the day on which the Bureau report is issued, when they are delivered to the Statistician by the Secretary or the Assistant Secretary. The combination for opening the safe in which such documents are kept is known only to the Secretary and the Assistant Secretary of Agriculture. Reports from special field agents and State statistical agents residing at points more than 500 miles from Washington are sent by telegraph, in cipher. Those in regard to speculative crops are addressed to the Secretary of Agriculture.

"Reports from the State statistical agents and special field service in relation to nonspeculative crops are sent in similar envelopes marked 'B' to the Bureau of Statistics and are kept securely in a safe until the data are required by the Statisticians in computing estimates regarding the crops to which they relate. The reports from the county correspondents, township correspondents, and other voluntary agents are sent to the Chief of the Bureau of Statistics by mail in sealed envelopes.

"The work of making the final crop estimates each month culminates at sessions of the Crop-Reporting Board, composed of five members, presided over by the Statistician and Chief of Bureau as chairman, whose services are brought into requisition each crop-reporting day from among the statisticians and officials of the

Bureau, and special field and State statistical agents who are called to Washington for the purpose.

"The personnel of the Board is changed each month. The meetings are held in the office of the Statistician, which is kept locked during sessions, no one being allowed to enter or leave the room or the Bureau, and all telephones being disconnected.

"When the Board has assembled, reports and telegrams regarding speculative crops from State and field agents, which have been placed unopened in a safe in the office of the Secretary of Agriculture, are delivered by the Secretary, opened, and tabulated; and the figures, by States, from the several classes of correspondents and agents relating to all crops dealt with are tabulated in convenient parallel columns; the Board is thus provided with several separate estimates covering each State and each separate crop, made independently by the respective classes of correspondents and agents of the Bureau, each reporting for a territory or geographical unit with which he is thoroughly familiar.

"With all these data before the Board, each individual member computes independently, on a separate sheet or final computation slip, his own estimate of the acreage, condition, or yield of each crop, or of the number, condition, etc., of farm animals for each State separately. These results are then compared and discussed by the Board under the supervision of the chairman, and the final figures for each State are decided upon.

"The estimates by States as finally determined by the Board are weighted by the acreage figures for the respective States, the result for the United States being a true weighted average for each subject. Thus, the figures for the United States are not straight averages, which would be secured by dividing the sum of the State averages by the number of States; but each State is given its due weight in proportion to its productive area for each crop.

"Reports in relation to cotton, after being prepared by the Crop-Reporting Board, and personally approved by the Secretary of Agriculture, are issued on the first or second day of each month during the growing season, and reports relating to the principal farm crops and live stock on the seventh or eighth day of each month. In order that the information contained in these reports may be made available simultaneously throughout the entire United States, they are handed, at an announced hour on report days, to all applicants and to the Western Union Telegraph Company and the Postal

Telegraph Cable Company, who have branch offices in the Department of Agriculture, for transmission to the Exchanges and to the press. These companies have reserved their lines at the designated time, and forward immediately the figures of most interest. A mimeograph or multigraph statement, also containing such estimates of condition or actual production, together with the corresponding estimates of former years for comparative purposes, is prepared and sent immediately to Exchanges, newspaper publications, and individuals. The same day printed cards containing the essential facts concerning the most important crops of the report are mailed to the 77,000 post-offices throughout the United States for public display, thus placing most valuable information within the farmer's immediate reach.

"Promptly after the issuing of the report, it, together with other statistical information of value to the farmer and the country at large, is published in the *Crop Reporter*, an eight-page publication of the Bureau of Statistics, under the authority of the Secretary of Agriculture. An edition of over 120,000 copies is distributed to the correspondents and other interested parties throughout the United States each month."

Technical Terms: Normal, Condition, Indicated Yield per Acre

To understand the official method of forecasting the size of agricultural crops, one must have clearly in mind the technical meaning of the terms normal, condition, and indicated yield per acre. The correspondents of the Bureau of Statistics of the Department of Agriculture are instructed to assume that a normal crop is to be represented by 100, and they are asked to express the condition of the crop in their respective districts, during successive months, as percentages of the normal. From these figures of condition supplied by its correspondents and agents, the Bureau of Statistics computes the indicated yield per acre for the several states, and for the whole country.

But what is a normal crop? Although the degree of efficiency of the crop-reporting service is largely dependent upon an accurate definition and sufficient understanding of this fundamental term, the Bureau of Statistics has been very slow to give an adequate description of its meaning. The official instruction which for a long time was given to the correspondents of the Department is here quoted at length:

The Normal. "So many of the reports of the Statistician of the Department of Agriculture are based upon a comparison with the 'normal' that it is a matter of the greatest importance that there should be a clear understanding of what the normal really means.

"To begin with, a normal condition is *not* an *average* condition, but a condition *above* the average, giving promise of *more than an average crop*.

"Furthermore, a normal condition does *not* indicate a *perfect* crop, or a crop that is or promises to be the very largest in quantity and the very best in quality that the region reported upon may be considered capable of producing. The normal indicates something *less* than this, and thus comes between the average and the possible maximum, being greater than the former and less than the latter.

"The normal may be described as a condition of perfect healthfulness, unimpaired by drought, hail, insects, or other injurious agency, and with such growth and development as may reasonably be looked for under these favorable conditions. As stated in the instruction to correspondents, it *does not* represent a crop of extraordinary character, such as may be produced here and there by the special effort of some highly skilled farmer with abundant means, or such as may be grown on a bit of land of extraordinary fertility, or even such as may be grown quite extensively once in a dozen years in a season that is extraordinarily favorable to the crop to be raised. A normal crop, in short, is neither deficient on the one hand nor extraordinarily heavy on the other. While a normal condition is but rarely reported for the entire corn, wheat, cotton, or other crop area, at the same time or in the same year, its local occurrence is by no means uncommon, and whenever it is found to exist, it should be indicated by the number 100.

"Sometimes a favorable season for planting is followed by a

favorable growing season, with no blight and no depredations by insects, the result being a normal condition. At other times the normal may be maintained by conditions that are exceptionally favorable in one or more particulars counterbalancing conditions that are unfavorable in other particulars. Thus, a crop may have had such an unusually good start that it may pass without injury through a period of drought that would otherwise have proved disastrous to it, or its more than ordinary vigor and potentiality may fully offset some slight injury from insects.

"The normal not being everywhere the same, in determining how near the condition of any given crop is to the normal, correspondents will usually find it an advantage to have a definite idea of what yield per acre would constitute a full normal crop in their respective districts; that is, how many bushels, pounds, or tons per acre of a particular crop would be produced in a season that was distinctly but not exceptionally favorable. In a region where 30 bushels of corn may be taken as the normal, a condition of 90 would give a prospect of a crop of 27 bushels, and 80 a crop of 24 bushels. If 40 bushels be considered the normal yield, 90 (or ten per cent less than the normal) would indicate a crop of 36 bushels, 80 one of 32 bushels, 70 one of 28 bushels.

"For the reason that the normal, represented by 100, does not indicate a perfect or the largest possible crop, it may occasionally be exceeded. The condition may be so exceptionally favorable as to promise a crop that will exceed the normal, and it will accordingly have to be expressed by 105, 110, or whatever other figures may seem warranted by the facts; 105 representing five per cent above the normal, 110 ten per cent, and so forth." [1]

The least that can be said about this definition of normal is that, as the individual farmer-correspondents must express the current condition of the crop as a percentage of the normal, the official Bureau leaves much to the individual farmers to determine. As late as August, 1916, a writer of great influence in the cotton trade has condemned the whole crop-reporting service because of the lack of precision in the instruc-

[1] *Crop Reporter*, May, 1899, p. 3. Cf. Circular 17 of the Bureau of Statistics, pp. 12–13.

tions that are sent to those who supply the primary statistical data:

"Those who report for the Crop Estimating Board are asked to make a mental comparison between existing conditions and an imaginary normal. They are instructed to assume that this indefinable normal is represented by 100 and to describe the present and its promise in figures that are supposed to be a percentage of an impossible perfection. The very difficulty of stating the theory upon which the reports are compiled shows how misleading they may be, but the practical impossibility of applying the theory utterly shatters any claims that such findings are entitled to scientific consideration." [1]

Although, as we have seen, the reports on the condition of the growing crops have been issued continuously since 1866, the Government authorities, until 1911, systematically refused to say what was to be inferred from their laboriously compiled tables. We have the repeated statement that "the Department, as is well known, makes no attempt to estimate in advance the probable yield of any agricultural product." [2] "The Department's reports previous to harvest" are intended "simply as a general epitome of the crop situation." [2] But the Department has been aware all along that the reports "are interpreted as furnishing a basis for quantitative forecasts of yield." [2] It is

[1] Theodore H. Price: "The Value and Defects of Government Crop Reports," *Commerce and Finance*, August 16, 1916, p. 915. Mr. Price's strictures with reference to the "indefinable normal" do not hold with the same degree of force since the improvement in the crop-reporting service in 1911. Whether the reports have any scientific value will be revealed in the course of this chapter.

[2] *Crop Reporter*, May, 1900, p. 6, and May, 1902, p. 4.

surely not amiss to say that the appropriations of public funds for the crop-reporting service have been made not because the Department has supplied "simply a general epitome of the crop situation," but because the claim has been urged that the crop-reporting service gives the public, and particularly the farmers, "early information concerning the supply" [1] of agricultural products.

Until the Department of Agriculture told us what its crop reports meant and how its elaborate tables as to crop condition were to be utilized to forecast the probable yield per acre, it was not possible to test the efficiency and utility of the crop-reporting service. The Department, since 1911, has given us the needed information, and before we pass to the actual testing of the degree of accuracy in the official forecasts we shall quote at length the official description of the "Interpretation of Crop Condition Figures."

"The Bureau of Statistics has this year for the first time given a quantitative interpretation to its monthly figures relating to the condition of growing crops; that is, has indicated the yield which the condition figures suggest. Much interest is manifested as to the method used in making such interpretations.

"It is assumed, in the first place, that average conditions at any time are indicative of average yields per acre; that conditions above an average at any time are indicative of yields above the average; and conditions below the average at any time are indicative of yields below the average. If at any time the condition of a growing crop is 5 per cent above the average condition for such time, it is assumed that the yield is more likely to be 5 per cent above the average yield than any other amount. If the condition at any time is 10 per cent below the average for such time, it is assumed that

[1] "Government Crop Reports: Their Value, Scope, and Preparation." U. S. Department of Agriculture. Bureau of Statistics. Circular 17, p. 7.

the yield is more likely to be 10 per cent below the average than any other amount.

"As a growing crop progresses toward maturity, its relation to an average condition is almost constantly changing; if the growing period becomes more favorable than the average, the prospects improve and the indicated yield enlarges; as the growing period becomes less favorable than the average, the prospect diminishes and the indicated yield lessens.

"In interpreting the condition figures it is necessary to determine what is an average condition at any time and what is the corresponding average yield. Different results will be obtained by using different bases. For instance, the condition of spring wheat on July 1 in the last 5 years (84.5 per cent of normal) averaged nearly 4 per cent lower than the average for the last 10 years (87.8); and the average yield per acre in the last 5 years (13.5 bushels) is about 2 per cent lower than the average for the last 10 years (13.8 bushels).

"The objection to a 10-year basis for determining either the average or normal yield of crops is that there is a gradual tendency of the average or normal yield per acre for the United States to increase from year to year, and therefore an average based upon a long series of years will be too low. For instance, the calculated equivalent of 100 condition for winter wheat at harvest time in the last 3 years averaged about 18.8 bushels, the average of the last 6 years was about 18.6 bushels, and for the last 10 years, 17.9 bushels.

"On the other hand, a 5-year basis includes so few years that one extreme or abnormal year in the 5 may so affect the average as to make it not representative of general average condition. For instance, the yield per acre of flaxseed in 1910 was abnormally low, 4.8 bushels, as compared with 9.4 in 1909, 9.6 in 1908, 9 in 1907, and 10.2 in 1906, the average of the 5 years being 8.6, which is lower than any year included in the average except 1910; the year 1910, therefore, ought to be omitted in obtaining a figure representing average conditions.

"After a study of the results obtained from using 5 years and 10 years, respectively, for basing an average, the advantage is found to be slightly in favor of the 5-year basis. In using the 5-year basis, however, it is proper to omit years of abnormal conditions.

"The process in the interpretation may be explained by an ex-

ample. The condition of corn on July 1, 1911, was 80.1 per cent of a normal condition; in the last 5 years the condition has averaged 85 per cent of a normal condition; thus the condition on July 1 is 5.8 per cent below the average condition (80.1 being 94.2 per cent of 85), and suggests a yield of 5.8 per cent below the average. In the last 5 years the yield averaged about 27.1 bushels; 94.2 per cent of 27.1 bushels (94.2 × 27.1) is nearly 25.5 bushels; therefore conditions are said to indicate a yield of 25.5 bushels. That is, if the condition of the corn crop be 5.8 per cent below the average at harvest time, a yield of 25.5 bushels is the most reasonable expectation; if less than the average adversity befall the crop before harvest, a larger yield may be expected; if more than the average adversity befall the crop, a yield less than 25.5 bushels may be expected.

"Another method of interpretation of the bureau's condition report has been used by some private statisticians, which may be explained here briefly by an example. The condition of the corn crop on July 1, 1911, is 80.1 per cent of normal; the average condition of the corn crop on October 1 for the last 5 years has been 80 per cent of a normal; thus the condition on July 1 (80.1) is 0.1 per cent above the average condition (80) on October 1 (the October report being the nearest to the harvest condition). The average yield being 27.1 bushels, 0.1 per cent above average would be nearly 27.1 bushels, the yield indicated on this basis, as against 25.5 bushels indicated by the method adopted by the Bureau of Statistics.

"The difference between the two methods is this: By the one adopted by the bureau it is assumed that from July 1 to harvest the average amount of variation will occur. (The 5-year average condition on July 1 is 85 per cent of a normal; the 5-year average condition on October 1 is 80.) By the other method it is assumed that no variation in condition will occur, notwithstanding that in the last 5 years the average change has been from 85 to 80, or a decline.

"The difference between the two methods of interpretation may also be shown as follows, using the same example: The average condition of corn July 1 is 85 and the average yield 27.1, hence the equivalent of 100 on July 1 is 31.9 bushels (27.1 × 100 ÷ 85); hence a condition of 80.1 on July 1 indicates a yield of 25.5 bushels (31.9 × 80.1 ÷ 100). This is the method adopted by the Bureau of Statistics.

"The other method referred to, used by some statisticians, is as follows: The average condition of corn on October 1 (the condition

report nearest to time of harvest) is 80 per cent, average yield, 27.1 bushels; hence the equivalent of 100 on October 1 (27.1 × 100 ÷ 80) is 33.9 bushels; this equivalent of 100 is used throughout the growing season as the equivalent of 100, and hence on July 1 when the condition is 80.1 it is interpreted as indicating a yield of 27.1 bushels (33.9 × 80.1 ÷ 100).

"The difference between the two methods is that the one adopted by the bureau allows for the natural variation as the season progresses, while the other does not.

"It may be pertinent to observe, considering the interpretation of crop condition figures, that the higher the condition of a crop the more sensitive it is; that is, liable to a decline before harvest. For example, of the last 10 years, the 5 which give the highest condition of winter wheat on May 1 averaged 91.8 per cent of normal, and the remaining 5 years of lowest condition on May 1 averaged 80.3. The 5 years which averaged 91.8 per cent on May 1 averaged 83.2 per cent on July 1, a drop of 8.6 points, or 9.4 per cent; the 5 years which averaged 80.3 per cent on May 1 averaged 79.6 on July 1, a drop of only 0.7 point, or 0.9 per cent. Neither method described takes into account this factor." [1]

The Accuracy of Forecasts Tested

If, for a moment, we review the description given by the Bureau of Statistics of its method of forecasting the probable yield per acre of a crop, we shall see that no reason is given for preferring the five years average over the ten years average as the basis of prediction, except that the former gives a "slightly better result." In what way the result of the five years method is better than the result of the ten years method is not stated, nor is there indicated any way by which we may determine how much better one method is than another. But it must be very clear that, for business and for scientific purposes, the measurement of the degree of

[1] *Crop Reporter*, July, 1911, pp. 53–55.

accuracy and reliability of forecasts is of the first importance.

Without a measure of the degree of accuracy and reliability of the forecasts the Crop-Reporting Board has no proper ground of choice between different methods of forecasting; farmers, brokers, and consumers are without adequate guidance in the planning of their enterprises; and scientific economists have no empirical tests of the degree of error in their analyses of the interrelation of phenomena. The Crop-Reporting Board, as we have seen, uses a five years average in preference to one of ten years. But would not a four years average, or a three years average, give better results than the five years average? The Bureau of Statistics issues annually six cotton reports, beginning with May. What is the value of these several reports as forecasts? Do the predictions become increasingly accurate with the approach of harvest? Are the reports for all of the months worthy of confidence, or are some of them so misleading as to suggest the wisdom of their discontinuance? If all of the forecasts are affected with error, is there a tendency of the error to favor the manufacturer or the farmer, and what are the risks assumed in planning one's enterprise according to the forecasts? The answers to all of these questions become possible when we have found an adequate method of measuring the degree of accuracy in the forecasts. To this problem we now turn.

The method of the Crop-Reporting Board in making the forecast of the yield per acre of cotton is, as we have seen, to assume that the ratio of the yield per acre for any given year to the mean yield per acre for

the preceding five years is equal to the ratio of the condition of any given month to the mean condition for the given month during the preceding five years. The method may be put in the form of symbols: Let Y = the yield per acre of cotton for the current year; $\bar{Y}_5$ = the mean yield per acre of cotton for the preceding five years; C = the condition of the crop for the current month; $\bar{C}_5$ the mean condition for the same month during the preceding five years. Then, according to the assumption of the Crop-Reporting Board, $Y/\bar{Y}_5 = C/\bar{C}_5$. We shall call $Y/\bar{Y}_5$ the yield-ratio; and $C/\bar{C}_5$, the condition-ratio.

In Table 5 we have the record, for a quarter of a century, of the actual yield per acre of cotton, and the yield as predicted, according to the method of the Crop-Reporting Board, from the condition of the crop on the 25th of September. The column marked $C/\bar{C}_5$ gives the ratio of the condition of the crop, at the end of September of any given year, to the mean condition of the crop, at the end of September, during the preceding five years. According to the method of the Crop-Reporting Board, the ratio $C/\bar{C}_5$ is the probable ratio that the yield per acre of any given year should bear to the mean yield of the preceding five years. But the column marked $Y/\bar{Y}_5$ gives the ratio of the actual yield per acre to the mean yield per acre of the preceding five years. An inspection of the Table shows that the two series — the predicted series and the actual series — are not equal. Is there any relation between the two series, and if so, how closely are they related, and what is the error made in using the condition-series to forecast the yield-series?

TABLE 5. — The Actual Yield Per Acre of Cotton in the United States Compared with the Yield as Forecast from the September Condition of the Crop

Year	Condition of crop September 25 C	Mean condition September 25, for the preceding 5 years $\bar{C}_5$	$C/\bar{C}_5$	Yield per acre of cotton in pounds of lint Y	Mean yield per acre of cotton, in pounds of lint, for the preceding 5 years $\bar{Y}_5$	$Y/\bar{Y}_5$
1885	78.0			163.9		
6	79.3			169.5		
7	76.5			182.8		
8	78.9			180.4		
9	81.5			177.0		
1890	80.0	78.8	101.5	187.0	174.7	106.8
1	75.7	79.2	95.6	179.4	179.3	100.1
2	73.3	78.5	93.4	209.2	181.3	115.4
3	70.7	77.9	90.8	148.8	186.6	79.7
4	82.7	76.2	108.5	191.7	180.3	106.3
5	65.1	76.5	85.1	155.6	183.2	84.9
6	60.7	73.5	82.6	124.1	176.9	70.2
7	70.0	70.5	99.3	181.9	165.9	109.6
8	75.4	69.8	108.0	219.0	160.4	136.5
9	62.4	70.8	88.1	184.1	174.5	105.5
1900	67.0	66.7	100.4	194.4	172.9	112.4
1	61.4	67.1	91.5	169.0	180.7	93.5
2	58.3	67.2	86.8	188.5	189.7	99.4
3	65.1	64.9	100.3	174.5	191.0	91.4
4	75.8	62.8	120.7	204.9	182.1	112.5
5	71.2	65.5	108.7	186.1	186.3	99.9
6	71.6	66.4	107.8	202.5	184.6	109.7
7	67.7	68.4	99.0	178.3	191.3	93.2
8	69.7	70.3	99.1	194.9	189.3	103.0
9	58.5	71.2	82.2	154.3	193.3	79.8
1910	65.9	67.7	97.3	170.7	183.2	93.2
11	71.1	66.7	106.6	207.7	180.1	115.3
12	69.6	66.6	104.5	190.9	181.2	105.4
13	64.1	67.0	95.7	182.0	183.7	99.1
14	73.5	65.8	111.7	207.9	181.1	114.8

By utilizing the methods described in the preceding chapter we not only may answer these questions, but we shall gain additional information of very great importance. Suppose we let the values in the $Y/\bar{Y}_5$ series be represented by y and the values in the $C/\bar{C}_5$ series be represented by x. Then, we know from the theory of correlation, if the association between the variables is linear, the closeness of their relation is measured by the coefficient of correlation r, and the straight line connecting the values of y with the values of x is described by the equation $(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$, where $\bar{y}$, $\bar{x}$ and σ_y, σ_x are the means and the standard deviations, respectively, of the y's and x's. By putting our reasoning into symbolic form we see the implied assumptions in the method of the Crop-Reporting Board. The method assumes that $y = x$, and, consequently, it implicitly assumes (1) that $r\frac{\sigma_y}{\sigma_x} = 1$, and (2) that $\bar{y} = \bar{x}$. The outcome of these assumptions we shall presently consider.

As the number of observations in this case is small, extending only through twenty-five years, the method of computing the coefficient of correlation differs slightly from the method that was illustrated in Chapter II; and for this reason, the process of computation is completely exemplified in Table 6. We find that the relation between the two variables y and x, for the month of September, is, $r = .685$.

The graph in Figure 8 shows the scatter diagram connecting the yield-series with the condition-series. The equation to the straight line connecting the two

TABLE 6. — CORRELATION OF THE SEPTEMBER CONDITION-RATIO AND THE YIELD-RATIO OF COTTON

Year	September Condition-Ratio $C/\bar{C}_5$	Yield-Ratio $Y/\bar{Y}_5$	Arbitrary Origin of Condition-Ratio at 100 x'	Arbitrary Origin of Yield-Ratio at 100 y'	$(x')^2$	$(y')^2$	$x'y'$	
							$+x'y'$	$-x'y'$
1890	101.5	106.8	1.5	6.8	2.25	46.24	10.20	
1	95.6	100.1	— 4.4	0.1	19.36	.01		.44
2	93.4	115.4	— 6.6	15.4	43.56	237.16		101.64
3	90.8	79.7	— 9.2	— 20.3	84.64	412.09	186.76	
4	108.5	106.3	8.5	6.3	72.25	39.69	53.55	
5	85.1	84.9	— 14.9	— 15.1	222.01	228.01	224.99	
6	82.6	70.2	— 17.4	— 29.8	302.76	888.04	518.52	
7	99.3	109.6	— 0.7	9.6	.49	92.16		6.72
8	108.0	136.5	8.0	36.5	64.00	1332.25	292.00	
9	88.1	105.5	— 11.9	5.5	141.61	30.25		65.45
1900	100.4	112.4	0.4	12.4	.16	153.76	4.96	
1	91.5	93.5	— 8.5	— 6.5	72.25	42.25	55.25	
2	86.8	99.4	— 13.2	— 0.6	174.24	.36	7.92	
3	100.3	91.4	0.3	— 8.6	.09	73.96		2.58
4	120.7	112.5	20.7	12.5	428.49	156.25	258.75	
5	108.7	99.9	8.7	— 0.1	75.69	.01		.87
6	107.8	109.7	7.8	9.7	60.84	94.09	75.66	
7	99.0	93.2	— 1.0	— 6.8	1.00	46.24	6.80	
8	99.1	103.0	— 0.9	3.0	.81	9.00		2.70
9	82.2	79.8	— 17.8	— 20.2	316.84	408.04	359.56	
1910	97.3	93.2	— 2.7	— 6.8	7.29	46.24	18.36	
11	106.6	115.3	6.6	15.3	43.56	234.09	100.98	
12	104.5	105.4	4.5	5.4	20.25	29.16	24.30	
13	95.7	99.1	— 4.3	— 0.9	18.49	.81	3.87	
14	111.7	114.8	11.7	14.8	136.89	219.04	173.16	
Totals			78.7 — 113.5 — 34.8	153.3 — 115.7 37.6	2309.72	4819.20	2375.59	180.40

Using the same symbols that we employed in Chapter II, we have

$$d_x = \frac{-34.8}{25} = -1.392; \quad \frac{\Sigma(x')^2}{25} = \frac{2309.72}{25} = 92.3928;$$

$$\sigma_x = \sqrt{\frac{\Sigma(x')^2}{25} - d_x^2} = 9.51; \quad \frac{\Sigma(x'y')}{25} = \frac{2375.59 - 180.40}{25} = 87.8076;$$

$$d_y = \frac{37.6}{25} = 1.504; \quad \frac{\Sigma(y')^2}{25} = \frac{4819.20}{25} = 192.7680;$$

$$\sigma_y = \sqrt{\frac{\Sigma(y')^2}{25} - d_y^2} = 13.80; \quad \frac{\Sigma(xy)}{25} = \frac{\Sigma(x'y')}{25} - d_x d_y = 89.9012;$$

$$r = \frac{\Sigma(xy)}{25\sigma_x\sigma_y} = .685.$$

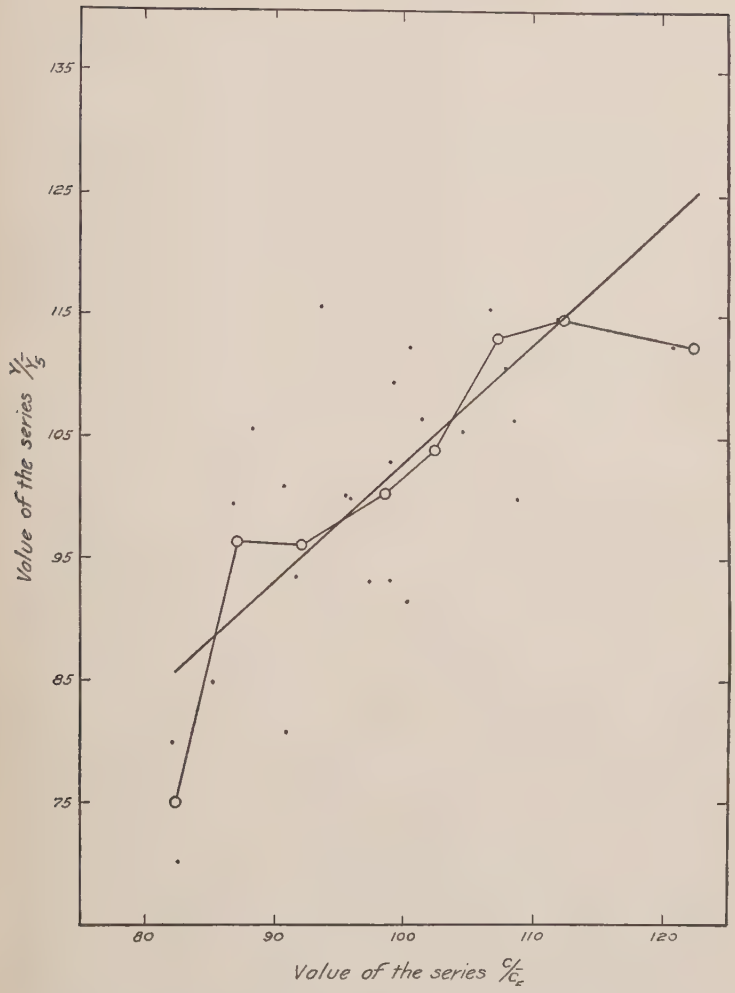

FIGURE 8. — Correlation of the actual yield-ratios of cotton with the forecasts from the September condition of the crop.

Equation to the straight line, $y = .994x + 3.49$, where y = the probable value of $Y/\bar{Y}_5$ and $x = C/\bar{C}_5$.

series, which is given on Figure 8, was computed from $(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$. For any given value of x, we may find the corresponding most probable value of y, either by substituting for x in the equation to the straight line and then solving for y, or, by using the graph of Figure 8 obtain the ordinate of the straight line corresponding to the given abscissa. Furthermore, we are able to measure the scatter of the observations about the straight line from the formula, $S = \sigma_y\sqrt{1 - r^2}$. In the particular case before us, $\sigma_y = 13.80$; $r = .685$; and, consequently, $S = 10.06$. In brief, we know the closeness of the relation between the two series from the value of $r = .685$; we know the law connecting the two series from the equation, $y = .99\,x + 3.49$; and we know the magnitude of the error made in using this law as a formula to predict the probable yield per acre, because $S = \sigma_y\sqrt{1 - r^2} = 10.06$.

If we examine the degree of accuracy of the forecast obtained by using the method of the Bureau of Statistics, we shall find that, for this month of September, the results agree very closely with the value of S which we have obtained by the method of correlation. The forecast-series in Table 5 is given in the column marked $C/\bar{C}_5$ and the actual series is given in the column marked $Y/\bar{Y}_5$. If we take the sum of the squares of the differences between $Y/\bar{Y}_5$ and $C/\bar{C}_5$, and divide by the number of cases, we shall have the mean square of the deviations of the actual series from the theoretical series, and this value is comparable with S^2. If we put S' equal to the square-root of the mean square of the deviations of the actual series from the predicted series, we may write,

$S' = \sqrt{\left\{\frac{\Sigma(Y/\bar{Y}_5 - C/\bar{C}_5)^2}{N}\right\}}$, where N is the number of years through which the series extends, and $\Sigma(Y/\bar{Y}_5 - C/\bar{C}_5)^2$ indicates the sum of the squares of the deviations, during the several years, of the actual series from the predicted series. The value of S' for the month of September is, $S' = 10.47$; and the value of S, we found just now, was $S = 10.06$. We see, accordingly, that for this month of September, the accuracy of the forecast by the method of the Bureau of Statistics is about as great as the accuracy of the method of correlation. A moment ago we found that the implicit assumptions in the official method are (1) that $r\frac{\sigma_y}{\sigma_x} = 1$, and (2) that $\bar{y} = \bar{x}$. For this particular month of September, $r = .685$; $\sigma_y = 13.80$; $\sigma_x = 9.51$; $\bar{y} = 101.5$; $\bar{x} = 98.6$. These values make $r\frac{\sigma_y}{\sigma_x} = .994$; and $\bar{y} - \bar{x} = 2.9$. The implicit assumptions in the official method are, therefore, in case of this one month of September, in close agreement with the facts, and, consequently, the results given by the two methods are almost the same.

Now that we have a means for testing the accuracy of the official method of forecasting we are in a position to answer the questions which we asked awhile ago, one of which was: What is the relative value of the official forecasts from the data for May, June, July, August, and September? In Table 7 we have a summary view of the important facts.

This Table reveals a number of facts of practical significance: (1) If we compare the values of S and S'

we find that, in case of all of the months, S is less than S'; that is to say, the method of correlation gives a more accurate result than the method of the Bureau of Statistics, although the difference between the accuracy of the two methods, except for the month of May, is very small. But this is the least important fact revealed by the Table.

TABLE 7. — DEGREES OF ACCURACY IN THE MONTHLY OFFICIAL FORECASTS OF THE YIELD PER ACRE OF COTTON IN THE UNITED STATES

Month	Relation between actual yield and predicted yield Value of r	$S = \sigma_y\sqrt{1-r^2}$	$S' = \sqrt{-\left\{\frac{\Sigma\left(Y/\bar{Y}_5 - C/\bar{C}_5\right)^2}{N}\right\}}$
May	—.049	13.79	16.05
June	.292	13.20	13.61
July	.595	11.09	11.51
August	.576	11.28	11.74
September	.685	10.06	10.47

(2) We see from the column giving the value of r that the May report as to condition and probable yield per acre has no value whatever; or, since it is supposed by farmers, brokers, and manufacturers to throw some light on the probable yield, we may put the case stronger and say that the report is worse than useless. The criticism is fortified by a consideration of the value of S' for the month of May, which is 16.05. This indicates that the root-mean-square of the deviations of the actual series from the predicted series exceeds the standard deviation of the actual series, which is $\sigma_y = 13.80$. It would, therefore,

be much more profitable to pay no heed whatever to the May report, which is issued about the first of June, and to assume that the mean value of the actual series for the past twenty-five years, namely, 101.5, will be the probable ratio of $Y/\bar{Y}_5$. For if one followed the official report on condition and prospective yield, one's error would be measured by $S' = 16.05$, whereas if no attention were paid to the report, the error would be $\sigma_y = 13.80$.

(3) The Table indicates further that the report for the month of June, which is issued about the first of July, has very little value. The correlation between the actual series and the predicted series is $r = .292$, and the value of $S = 13.20$, while $S' = 13.61$. Since the value of σ_y is 13.80, we are justified in saying that the report for June has very little, if any, value as a basis for predicting the probable yield.

(4) Another fault revealed by the Table is that the July forecast is at least as accurate as the August forecast. This is shown by the value of r for July exceeding the value of r for August, and by the values of S and S' for July being less than the corresponding values for August. As far as these two months are concerned it is not true that as the harvest approaches the reports are increasingly accurate.

Our method of testing the accuracy of the official forecasts brings out another fault of practical importance and of theoretical interest. We have agreed to measure the scatter of the actual yield-ratio about the predicted yield-ratio by $S' = \sqrt{\left\{\frac{\Sigma(Y/\bar{Y}_5 - C/\bar{C}_5)^2}{N}\right\}}$. As in computing this value one subtracts for each year

$C/\bar{C}_5$ from $Y/\bar{Y}_5$, it is possible to see how often the predicted yield-ratio exceeds or falls short of the actual yield-ratio. The question is of importance because, if the forecasts show a tendency to fall short of the actual yield, the effect of the forecasts will be to create an undue rise in price and thereby favor the producers. The opposite result would be the case if the forecasts show a tendency to exceed the actual yield. The record of twenty-five years — in Table 8 — shows that there has been a tendency of the forecasts of each month to favor the producer. In the report of the condition of the crop for the month of May, the official method gives a forecast falling short of the actual yield-ratio, 19 times out of 25; for the month of June, 16 times out of 25; for the month of July, 15 times out of 25; for August, 16 times out of 25; for September, 15 times out of 25. These figures undoubtedly show a tendency in the official method of forecasting to give an underestimate of the probable yield.

The cause of this bias in the method does not lie on the surface. It might be supposed that the cause is due to the tendency of the farmer correspondents of the Department of Agriculture to take a pessimistic view of the agricultural outlook. But such an imputation of pessimism to the farmer is not warranted by the experience of the Bureau of Statistics. In the *Crop Reporter*, the official "medium of communication between the Division of Statistics and the crop reporters of the Department of Agriculture," we are told:

". . . Correspondents are frequently reminded" that "there has always been a tendency to overestimate the average yield per acre. This is accounted

TABLE 8. — The Tendency of the Official Method of Forecasting to Underestimate the Yield Per Acre of Cotton in the United States

Year	Value of $\left(Y/\bar{Y}_5 - C/\bar{C}_5\right)$				
	May	June	July	August	September
1890	+ 8.6	+ 6.0	+ 6.8	+ 5.6	+ 5.3
1	+ 4.7	+ 1.3	— 0.8	+ 1.9	+ 4.5
2	+ 19.1	+ 19.1	+ 23.6	+ 24.3	+ 22.0
3	— 18.7	— 14.1	— 12.2	— 8.6	— 11.1
4	+ 4.2	+ 3.8	— 0.3	+ 0.3	— 2.2
5	— 8.3	— 8.8	— 5.1	— 2.6	— 0.2
6	— 43.8	— 37.4	— 24.8	— 12.2	— 12.4
7	+ 14.3	+ 10.5	+ 4.3	+ 4.1	+ 10.3
8	+ 34.3	+ 31.2	+ 27.1	+ 29.4	+ 28.5
9	+ 7.9	+ 6.1	+ 7.4	+ 15.1	+ 17.4
1900	+ 17.9	+ 26.3	+ 21.9	+ 18.1	+ 12.0
1	+ 0.5	0.0	+ 1.2	— 5.9	+ 2.0
2	— 13.3	— 1.0	+ 0.8	+ 12.0	+ 12.6
3	+ 6.0	— 0.3	— 5.7	— 23.9	— 8.9
4	+ 13.4	+ 4.3	— 2.3	— 6.5	— 8.2
5	+ 7.1	+ 5.2	+ 7.8	+ 2.2	— 8.8
6	+ 6.8	+ 7.6	+ 7.5	+ 6.1	+ 1.9
7	+ 8.1	+ 5.4	+ 2.0	— 2.8	— 5.8
8	+ 0.7	+ 1.0	+ 0.3	+ 4.8	+ 3.9
9	— 22.9	— 13.1	— 8.4	— 3.5	— 2.4
1910	— 11.1	— 10.8	— 4.2	— 6.4	— 4.1
11	+ 5.0	+ 2.8	+ 0.6	+ 14.2	+ 8.7
12	+ 7.0	+ 4.0	+ 8.4	+ 0.9	+ 0.9
13	+ 2.5	— 1.9	— 1.4	+ 4.4	+ 3.4
14	+ 24.0	+ 16.6	+ 17.5	+ 4.0	+ 3.1

for by local pride, by the publicity that is given to large individual yields, and by forgetfulness of the fact that there is in every agricultural community a large number of farms on which, during even the most favor-

able seasons, the yield of the particular crop is small." [1] (*Crop Reporter,* November, 1899, p. 1.)

Moreover, even if the farmer could be proved to take either a pessimistic or an optimistic view of the outcome of the crops, the effects of his bias would be eliminated in the official method of forecasting. The official formula is $Y/\bar{Y}_5 = C/\bar{C}_5$, and the numerator and the denominator in the fraction $C/\bar{C}_5$ would be equally affected by the farmer's bias; and, consequently, its influence would be eliminated in the ratio. The same thing would be true if there were a tendency in the Crop-Reporting Board either to overestimate or to underestimate the condition of the growing crop.

The fault lies with the method of forecasting and not with the farmers, or with the Crop-Reporting Board. Let us recall what was said awhile ago about the implicit assumptions in the official method. The official formula is $Y/\bar{Y}_5 = C/\bar{C}_5$, and we have written this in the form $y = x$, where $y = Y/\bar{Y}_5$, and $x = C/\bar{C}_5$. The official formula assumes that the relation between y and x is linear, but we know that the general formula

[1] The purport of the above quotation seems to be contradicted by a later official statement. In the *Crop Reporter* for January, 1905, there is an account of a "Hearing Before the Committee of Agriculture of the House of Representatives," relating to the methods of estimating the acreage, production, and yield per acre of cotton. Mr. John Hyde, at that time Chief of the Bureau of Statistics of the Department of Agriculture, testified, in part, as follows:

"The Chairman. Right there, let me ask you a question. Do you find that your sources of information, as a rule, are prone to underestimate?

"Mr. Hyde. They always have been and, as a consequence, the Department has never overestimated a crop.

"The Chairman. Your correspondents that bring the information in to you are prone to underestimate?

"Mr. Hyde. With very few exceptions; yes, sir."

for a linear relation between y and x is $(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$. Since the official method assumes that $y = x$, it implicitly assumes (1) that $r \frac{\sigma_y}{\sigma_x} = 1$, and (2) $\bar{y} - \bar{x} = 0$. It takes no account whatever of the degree of correlation between the two series, nor of the relative variabilities of the two series, nor of the difference in the mean values of the two series; and here lies the explanation of the tendency of the official method to give an underestimate of the yield per acre of cotton.

TABLE 9. — The Implicit Assumptions in the Official Method of Forecasting Are: (1) $r \frac{\sigma_y}{\sigma_x} = 1$; (2) $\bar{y} - \bar{x} = 0$

Month	r	σ_y	σ_x	$r \frac{\sigma_y}{\sigma_x}$	$\bar{y}$	$\bar{x}$	$\bar{y} - \bar{x}$
May	—.049	13.80	6.45	—.105	101.5	98.5	3.0
June	.292	13.80	6.15	.655	101.5	99.0	2.5
July	.595	13.80	7.17	1.145	101.5	98.6	2.9
August	.576	13.80	9.18	.866	101.5	98.5	3.0
September	.685	13.80	9.51	.994	101.5	98.6	2.9

The accompanying Table 9 shows how very far from agreement with the facts are the implicit assumptions in the official method. For all five of the months between seeding and harvest, the mean of the actual yield-ratio exceeds the mean of the predicted yield-ratio; that is to say, $\bar{y}$ is greater than $\bar{x}$; and in none of the months, except September, is the value of $r \frac{\sigma_y}{\sigma_x}$ ap-

proximately equal to unity. The figures for September afford the simplest illustration of the inherent bias in the official method. Since $r\frac{\sigma_y}{\sigma_x}$ is approximately equal to unity for the month of September, the general equation to the line connecting y with x, namely, $(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$, may be written $(y - \bar{y}) = (x - \bar{x})$; or, transposing the $\bar{y}$, we may write the equation in the form $y = x + (\bar{y} - \bar{x})$. The official method assumes that $(\bar{y} - \bar{x})$ is equal to zero, when, according to the actual figures for the month of September, $(\bar{y} - \bar{x}) = 2.9$. The equation to the straight line ought therefore to be $y = x + 2.9$, whereas the official formula is $y = x$, and, consequently, it gives a forecast of the value of y which, on the average, falls short of the true value by 2.9.

One of the questions naturally suggested with reference to the official method of forecasting is whether a four years progressive average, or a three years progressive average, would not give as good results, when used as a basis of forecasting, as the five years progressive average that is adopted by the Bureau of Statistics. The device that we have used to test the accuracy of the official forecasts enables us to obtain definite information. The results of the necessary computations are given in Tables 10 and 11.

The findings in Table 10 seem to justify the following conclusions:

(1) A comparison of the values of S and S', in each of the main divisions of the Table, shows that, in all of the fifteen cases, the value of S is less than the value

TABLE 10. — Results of Forecasting the Yield Per Acre of Cotton by the Method of Progressive Averages

Month	Progressive Average of Five Years		
	Relation between actual yield and forecast yield. Value of r	Scatter about the line of regression $S = \sigma_y\sqrt{1-r^2}$. $\sigma_y = 13.80$	Scatter about the line of progressive averages $S' = \sqrt{\frac{\Sigma\left(\frac{Y}{\bar{Y}_5} - \frac{C}{\bar{C}_5}\right)^2}{N}}$
May	—.049	13.79	16.05
June	.292	13.20	13.61
July	.595	11.09	11.51
August	.576	11.28	11.74
September	.685	10.06	10.47

Month	Progressive Average of Four Years		
	Relation between actual yield and forecast yield. Value of r	Scatter about the line of regression $S = \sigma_y\sqrt{1-r^2}$. $\sigma_y = 13.84$	Scatter about the line of progressive averages $S' = \sqrt{\frac{\Sigma\left(\frac{Y}{\bar{Y}_4} - \frac{C}{\bar{C}_4}\right)^2}{N}}$
May	—.073	13.80	16.12
June	.271	13.32	13.45
July	.576	11.31	11.58
August	.564	11.43	11.77
September	.683	10.11	10.42

Month	Progressive Average of Three Years		
	Relation between actual yield and forecast yield. Value of r	Scatter about the line of regression $S = \sigma_y\sqrt{1-r^2}$. $\sigma_y = 14.54$	Scatter about the line of progressive averages $S' = \sqrt{\frac{\Sigma\left(\frac{Y}{\bar{Y}_3} - \frac{C}{\bar{C}_3}\right)^2}{N}}$
May	.059	14.51	16.04
June	.380	13.45	13.61
July	.650	11.03	11.35
August	.599	11.64	11.86
September	.724	10.03	10.26

of S'; but that the difference is always small and theoretically insignificant, except for the month of May. For that one month, whether one employs the official formula with a five, four, or three years progressive mean, the value of S', since in each case it exceeds the corresponding value of σ_y, shows that the forecast is worse than useless;

(2) When attention is paid to the probable errors of S and S', there is really no difference in the accuracy of the forecasts whether they are based upon a five, four, or three years progressive average;

The findings in Table 11 enable us to add to the foregoing,

(3) The disposition of the official formula to give an underestimate of the yield per acre is shown, no matter whether the progressive average is one of three, four, or five years; but the bias is perhaps less in case of the three years average than in case of the five years average;

(4) The correlation prediction formula shows no bias.

Our general conclusion is that because of its greater accuracy and freedom from bias the correlation formula, with either a three or five years progressive average, is preferable to the official formula with the five years progressive average.

Acreage and Production

The two factors that are used to estimate the probable size of the cotton crop are the probable yield per acre and the number of acres under cultivation. The problem of forecasting the probable yield per acre has

TABLE 11. — The Number of Times in Twenty-five Years that the Actual Yield-Ratio Exceeded the Predicted Yield-Ratio. Official Formula $Y/\bar{Y} = C/\bar{C}$; Correlation Formula $(y-\bar{y}) = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$

Month	Five years progressive means			Four years progressive means			Three years progressive means		
	Number of times actual yield-ratio exceeded the predicted yield-ratio		Correlation formula where $x = C/\bar{C}_5$ and y = the predicted yield-ratio	Number of times actual yield-ratio exceeded the predicted yield-ratio		Correlation formula where $x = C/\bar{C}_4$ and y = the predicted yield-ratio	Number of times actual yield-ratio exceeded the predicted yield-ratio		Correlation formula where $x = C/\bar{C}_3$ and y = the predicted yield-ratio
	Official formula	Correlation formula		Official formula	Correlation formula		Official formula	Correlation formula	
May	19	13	$y = -.105x + 111.84$	17	12	$y = -.146x + 115.84$	17	13	$y = .117x + 89.66$
June	16	13	$y = .655x + 36.67$	16	13	$y = .596x + 42.24$	14	13	$y = .867x + 15.10$
July	15	11	$y = 1.145x - 11.44$	12	10	$y = 1.061x - 3.60$	12	10	$y = 1\ 177x - 15.31$
August	16	12	$y = .866x + 16.21$	15	12	$y = .862x + 16.14$	14	12	$y = .92x + 10.04$
September	15	12	$y = .994x + 3.49$	13	11	$y = 1.02x + .55$	13	12	$y = 1.066x - 4.48$

already been dealt with, and only a few words need be said about the official method of estimating acreage.

Throughout the whole period under investigation, 1890–1914, the Bureau of Statistics has again and again warned the public that its figures referring to acreage are merely estimates and not the results of extensive measurements such as are used by the Bureau of the Census. It has issued its reports as "the best available data, representing the fullest information obtainable at the time they are made," [1] and it has frankly pointed out the limitations of the method which it has felt compelled to follow in estimating acreage.

The census figures of the acreage devoted to the several crops, which have appeared every ten years, have been taken by the Bureau of Statistics as the foundation upon which to base its calculation as to the acreage under cultivation during the intercensal years. For each year between the census surveys, correspondents were asked to observe whether, as compared with the preceding year, there had been an increase or a decrease in the acreage of cotton in their respective districts, and to express the change as a percentage change. The Bureau of Statistics, using the last returns of the Census as the best available data, has computed the absolute value of the combined percentages of its correspondents and has issued the result as the Department's estimate of the acreage of the cotton crop for the current year. The Bureau has regarded each of its estimates merely as "a consensus of

[1] Annual Report of the Bureau of Statistics for the Fiscal Year 1911–1912. *Crop Reporter*, December, 1912.

judgment of many thousands of correspondents,"[1] and it has pointed out "that estimates made monthly from year to year, following each other during a period of 10 years, without means of verification or correction, are likely to be more or less out of line with conditions at the end of the 10-year period as disclosed by actual census enumerations. Cumulative errors, impossible of discovery, are likely to occur and cannot be corrected until census reports are available."[2] At the appearance of new census figures the Bureau of Statistics has revised its estimates of the preceding intercensal years, and the more recent census figures have been used as the basis of estimates for the following years.

It would doubtless be possible to test the degree of accuracy in the method employed by the Bureau of Statistics for calculating the acreage of the crops; or, to be more exact, it would be possible to test how nearly the preliminary estimates correspond with the revised estimates. As far as I am aware this test has never been carried out. The Bureau reports that, in case of some of the crops, there has been a considerable difference in the two estimates.[3] If the test were made and the method were found to be unsatisfactory, the problem would then present itself of finding a better method, and the solution of the problem would be sought in either of two directions: Either the direct measurement, such as is used by the Bureau of the Census, must

[1] Annual Report of the Bureau of Statistics for the Fiscal Year 1911–1912.

[2] *Ibidem.*

[3] *Crop Reporter*, May, 1900, p. 2. "Department of Agriculture and he Census."

be applied more frequently than ten years, and the method of estimates employed by the Bureau of Statistics be checked up at shorter intervals; or else the quantitative connections between variations in acreage and the variations in other economic factors must be discovered, and the acreage be then computed from these known connections. The former solution, which is undoubtedly the best, is urged by the Department of Agriculture.[1] But an agricultural survey is extremely expensive, and its results are frequently made known when, for many practical purposes, it is too late. The Bureau of Statistics has reported that "the results of the agricultural census which related to 1909 were not published in time to permit a revision of estimates of this Bureau until the close of 1911." [2] Furthermore, while the Bureau of Statistics makes its estimates for current use, the estimate of the acreage of cotton is not published until about July 1. But there are a number of industries dependent upon the acreage of cotton which would profit by having a reliable estimate earlier in the year. Would it not be possible to have a fair estimate of the probable acreage even before the crop is planted?

In Table 12 there is an illustration of a method by which a solution may be obtained of the problem that has just been described. The acreage planted in cotton, any given year, is largely dependent upon what has been the fortune, good or bad, of the cotton farmers in preceding years. If, for example, the price of cotton has been falling, few acres will be seeded in that particular

[1] Annual Report of the Bureau of Statistics for the Fiscal Year 1911–1912.

[2] *Ibidem.*

TABLE 12. — Percentage Change in the Acreage of Cotton and Percentage Change in the Production of Cotton Lint

Year	Acreage of Cotton (Thousands of acres)	Absolute change in acreage	Percentage change in acreage	Production of cotton lint (Millions of bales)	Absolute change in production	Percentage change in production
1888				6.92		
9	20,180			7.47	+ 0.55	+ 7.95
1890	21,886	+ 1706	+ 8.45	8.56	+ 1.09	+ 14.59
1	23,876	+ 1990	+ 9.09	8.94	+ 0.38	+ 4.44
2	15,228	— 8648	— 36.22	6.66	— 2.28	— 25.50
3	23,837	+ 8609	+ 56.53	7.43	+ 0.77	+ 11.56
4	24,959	+ 1122	+ 4.71	10.03	+ 2.60	+ 34.99
5	21,896	— 3063	— 12.27	7.15	— 2.88	— 28.71
6	32,823	+ 10927	+ 49.90	8.52	+ 1.37	+ 19.16
7	28,861	— 3962	— 12.08	10.99	+ 2.47	+ 28.99
8	25,174	— 3687	— 12.78	11.44	+ 0.45	+ 4.10
9	24,278	— 896	— 3.56	9.35	— 2.09	— 18.27
1900	24,982	+ 704	+ 2.90	10.12	+ 0.77	+ 8.24
1	26,897	+ 1915	+ 7.67	9.51	— 0.61	— 6.03
2	26,940	+ 43	+ 0.16	10.63	+ 1.12	+ 11.78
3	26,952	+ 12	+ 0.04	9.85	— 0.78	— 7.34
4	31,350	+ 4398	+ 16.32	13.44	+ 3.79	+ 36.45
5	27,205	— 4145	— 13.22	10.58	— 2.86	— 21.28
6	31,301	+ 4096	+ 15.06	13.27	+ 2.69	+ 25.43
7	29,848	— 1453	— 4.64	11.11	— 2.16	— 16.28
8	32,493	+ 2645	+ 8.86	13.24	+ 2.13	+ 19.17
9	31,060	— 1433	— 4.41	10.00	— 3.24	— 24.47
1910	32,467	+ 1407	+ 4.53	11.61	+ 1.61	+ 16.10
11	36,045	+ 3578	+ 11.02	15.69	+ 4.08	+ 35.14
12	34,283	— 1762	— 4.89	13.70	— 1.99	— 12.68
13	37,089	+ 2806	+ 8.18			

crop. There should therefore, in normal times, be some relation between the percentage change in the price of cotton last year over the preceding year and the percentage change in the acreage of cotton this year over

last year. In Table 12 the data are presented with which to compute the relation between the two variables, namely, the percentage change in the acreage of a given year over the acreage of the preceding year, and the percentage change in the price of cotton from the price prevailing two years before the current year to the price the year before the current year. In the same way that this correlation Table was prepared, similar Tables were compiled connecting the percentage change in the acreage of cotton with the percentage change, in the preceding year, of other variables. A summary of the calculations is here given:

The correlation between the percentage change in the acreage of cotton and

(1) the percentage change of the year before in the total production of cotton lint, $r = -.641$;

(2) the percentage change of the year before in the price per pound of cotton lint, $r = .532$;

(3) the percentage change of the year before in the value of the yield per acre of cotton lint, $r = .508$;

(4) the percentage change of the year before in the acreage of cotton, $r = -.492$;

(5) the percentage change of the year before in the yield per acre of cotton, $r = -.217$;

(6) the percentage change of the year before in the index number of general wholesale prices, $r = .005$.

From these calculations it is clear that even before the cotton crop is planted, it is possible to forecast the probable acreage with substantially the same degree of accuracy with which the Bureau

of Statistics can forecast the yield per acre of cotton at the first of September. We know from the results of the preceding chapter that when the correlation between two variables is linear, the scatter of the observations about the line of regression is measured by $S = \sigma_y \sqrt{1 - r^2}$. The degree of accuracy with which we can forecast results is, therefore, dependent upon the two factors σ_y and $\sqrt{1 - r^2}$. If we make allowance for the difference between the values of σ_y in case of two series, the relative degree of accuracy with which we can forecast results is dependent upon $\sqrt{1 - r}$; the smaller the value of $\sqrt{1 - r}$, the better the forecast. We have found that the correlation between the actual yield of cotton and the yield as predicted by means of the official formula is, at the first of August, $r = .595$, and at the first of September, $r = .576$. The correlation between the percentage change in the acreage of cotton for any given year and the percentage change in the production of cotton the preceding year is $r = -.641$; and the correlation between the percentage change in the acreage of any given year and the percentage change in the price per pound of cotton lint the preceding year is $r = .532$. It is true, therefore, that when allowance is made for the difference in the variabilities of the things compared, it is possible to forecast the acreage of cotton several months before the crop is planted, with as great a degree of accuracy as the Bureau of Statistics can forecast the possible yield per acre of cotton at the first of September.[1]

[1] It will be understood, I hope, that I am not offering this method of forecasting acreage as a substitute for the method employed by the Bureau of Statistics. I present it for its practical value in supplement-

If we gather into a summary the principal points that have been made in this chapter, we may list them as follows:

1. By means of a vast and remarkable statistical organization with thousands of correspondents, paid and unpaid, the Department of Agriculture collects its data referring to the condition of the cotton crop, and issues monthly reports during the period between seed-time and harvest. As the reports have great influence upon the price of cotton, every precaution is taken to prevent any leakage of information before the final conclusions are given to the public.

2. The Bureau of Statistics of the Department of Agriculture has devised a method of forecasting, from the monthly records of the condition of the crop, the probable yield per acre of cotton. As, however, the raw data are largely estimates supplied by its correspondents, the Bureau has regarded the material upon which its estimates are based as a "consensus of the opinion of the well-informed." Although the figures descriptive of the condition of the crop are expressed in percentages and decimals of percentages, the Bureau is aware of the insecure foundation upon which its forecasts rest. Nevertheless the forecasts have far-reaching effects. Indeed, as the Cotton Belt of the United States produces about 75 per cent of the world's cotton crop, the degree of accuracy in the work of the Bureau of Statistics of our Department of Agriculture produces its effect throughout the civilized world.

3. When the cotton reports are subjected to a critical

ing the work of the Bureau of Statistics, and for its theoretical value in illustrating that our economic activity is such a matter of routine that it admits of prediction.

examination as to the extent of their reliability and the degree of accuracy of the method of forecasting the probable yield per acre, the chief facts discovered are these:

(a) The May report – that is, the report referring to the condition of the crop at the end of May — has no value whatever as a basis upon which to forecast the average yield per acre of cotton. The percentages referring to the condition of the crop for the whole cotton section are arithmetical means of wild guesses; any forecasts based upon the May figures are spurious; and any money that changes hands in consequence of the forecasts are losses and gains resulting from a simple gamble;

(b) The June report — which is issued about the first of July — as far as it refers to the average condition of the cotton crop in the whole country, has a measurable, but small, value as a basis of forecasting the ultimate average yield per acre. When, for a period extending over a quarter of a century, the degree of relation between the predicted yield and the actual yield is properly measured, it is found that the coefficient of correlation between the two series, the forecast series and the actual series, is $r = .292$;

(c) The July, August, and September reports have a decided value as bases of forecasting the average yield per acre. The coefficients measuring the closeness of the relation between the predicted series and the actual series are, for July, $r = .595$; for August, $r = .576$; for September, $r = .685$. In this chapter and in chapter II, methods are described by which the degree of reliability of these reports as bases of forecasting the ultimate yield per acre are measured;

(d) The method of the Bureau of Statistics by which, from the reports on the condition of the crop, forecasts are made as to the ultimate yield per acre, has inherent defects that lead to an underestimate of the yield per acre, and thereby favors the producers. The official method of forecasting, if applied to the data referring to the condition of the crop, during a period of twenty-five years, gives a predicted yield per acre which, out of a total of 25 years, is an underestimate 19 times when based upon the May condition; 16 times when based upon the June condition; 15 times, in case of July; 16, in case of August; and 15, in case of September;

(e) It is possible to construct a better prediction formula than the one used by the Bureau of Statistics — a formula more accurate in its forecasts and entirely free from any disposition either to underestimate or to overestimate the yield per acre.

4. The official estimate of the acreage planted in cotton is published about July 1, and the two factors in estimating the ultimate production are the acreage and the yield per acre. By means of a method described in this chapter, it is possible to forecast the acreage, several months before the crop is planted, with a degree of precision as great as that of the official forecast of the yield, after the crop has completed its growth and is about to be harvested.

CHAPTER IV

FORECASTING THE YIELD OF COTTON FROM WEATHER REPORTS

"The essence of science . . . consists in inferring antecedent conditions, and anticipating future evolutions, from phenomena which have actually come under observation."

— Lord Kelvin.

In the preceding chapter the methods and results of the Department of Agriculture were examined with reference to the accuracy of the method of forecasting and the degree of value of the crop reports as forecasts of the yield per acre of cotton. The inquiry was concerned with the official statistics bearing upon the monthly condition of the growing crop and upon the annual yield per acre of cotton in the whole Cotton Belt. Holding the results of this inquiry in mind, we shall consider, in the present chapter, the possibility of forecasting the yield per acre of cotton simply from the current reports of the Weather Bureau as to rainfall and temperature in the several cotton states. Inasmuch as the investigation entails in case of each state a considerable amount of statistical computation, the inquiry will not be extended to the whole area of the Cotton Belt, but will be limited to a few representative states. Remembering always that our conclusions are based upon the detailed investigation of the representative states we shall maintain these theses:

(1) That some of the official reports referring to the individual states are valuable as forecasts, but that

others are worse than useless in the sense of supplying erroneous instruction as to the crop outlook, and thereby suggesting a misdirection of activity on the part of farmers, dealers, and manufacturers;

(2) That even in case of the useful forecasts, the official method does not extract the full amount of truth contained in the laboriously collected data;

(3) That notwithstanding the vast official organization for collecting and reducing data bearing upon the condition of the growing crop, it is possible, by means of mathematical methods, to make more accurate forecasts than the official reports, in the matter of the prospective yield per acre of cotton, simply from the data supplied by the Weather Bureau as to the current records of rainfall and temperature in the respective cotton states;

(4) That the principles and methods of forecasting the yield of cotton may be utilized in forecasting the yield of other agricultural crops.[1]

The Official Forecasts of the Yield of Representative States

In 1914 the total production of cotton in the United States was 16,135,000 bales of 500 pounds gross weight. This total product, with the exception of a negligibly small amount produced in other parts of the country, was the yield of thirteen states in the Cotton Belt. The four states of largest yield were Texas, with its 4,592,000 bales; Georgia, with 2,718,000 bales; Alabama, with

[1] This thesis will not be developed in the present Essay. The probability of its truth is suggested by the researches in Chapter III of *Economic Cycles: Their Law and Cause.*

1,751,000 bales; South Carolina, with 1,534,000 bales. In all, these four states produced 10,595,000 bales or sixty-five per cent of the whole crop.[1] Furthermore, Texas, with its enormous yield, is representative of the conditions of production in the extreme Southwest; Georgia and South Carolina exemplify the conditions at the other extreme of the Cotton Belt on the Atlantic Coast; Alabama typifies the conditions on the Gulf of Mexico. These four states, which are in direct order of their importance, happen to illustrate the weather conditions in the cotton-growing section and are, for the purpose of our inquiry, representative of the conditions of production in the whole of the Cotton Belt.

In order to measure the degree of accuracy of the official method of forecasting the yield per acre of cotton in these representative states, we shall recall the description of the forecasting formula which was given in the preceding chapter, together with the description of the coefficient measuring the degree of accuracy of the formula when it is actually applied to the given data. In symbolic terms the forecasting formula is $C/\bar{C}_5 = Y/\bar{Y}_5$, where C is the condition of the cotton crop in the current month; $\bar{C}_5$ is the mean condition of the crop in the same month during the five years preceding the given year; Y is the prospective yield per acre of the present year; and $\bar{Y}_5$ is the mean yield per acre during the preceding five years. The value of $C/\bar{C}_5$ is called the condition-ratio, or the forecasting ratio, and $Y/\bar{Y}_5$ is called the yield-ratio. We make a distinction between the theoretical yield-ratio and the actual yield-ratio. When in $Y/\bar{Y}_5$, Y is the predicted value of the yield per

[1] *Yearbook of the Department of Agriculture*, 1915, p. 470.

acre, then $Y/\bar{Y}_5$ is called the theoretical yield-ratio; and when Y is the actual value of the yield per acre, $Y/\bar{Y}_5$ is called the actual yield-ratio. Theoretically, if the forecasting formula were perfectly accurate, the actual yield-ratio $Y/\bar{Y}_5$ should always be equal to $C/\bar{C}_5$, but, as a matter of record, the actual yield-ratios differ from the condition-ratios, and a measure of the degree of accuracy of the forecasting formula, must, obviously, be some function of the difference between the actual yield-ratios and the theoretical yield-ratios; that is to say, the coefficient measuring the accuracy of the official formula must be some function of $(Y/\bar{Y}_5 - C/\bar{C}_5)$. For reasons that are fully set forth in the preceding chapter we have taken as the coefficient measuring the degree of accuracy of the official formula the value of S' where $S' = \sqrt{\left\{\frac{\Sigma(Y/\bar{Y}_5 - C/\bar{C}_5)^2}{N}\right\}}$, and N equals the number of observations.

With this understanding of the formula which we shall employ, we shall now give the results of the test of the degree of accuracy with which the official forecasting formula would enable one to predict the actual yield ratio during the 21 years from 1894 to 1914. The results are collected in the accompanying Table 13.

In the preceding chapter we showed why the official formula has increasing accuracy as the value of S' becomes smaller and smaller. The object of any forecasting of phenomena is, of course, to get as accurate an appreciation as possible of the phenomena.

S' represents the degree of approximation of the actual yield-ratios to the predicted yield-ratios, and σ_y gives the variability of the actual yield-ratios for the entire

period covered in the investigation, which, in this case, is the 21 years from 1894 to 1914.

TABLE 13. — TESTS OF THE ACCURACY OF THE OFFICIAL FORMULA THAT IS USED IN FORECASTING THE PROBABLE YIELD PER ACRE OF COTTON FROM THE MONTHLY CONDITION OF THE CROP

The Representative States	The Standard Deviation of the Yield Ratio σ_y	The Accuracy of the Forecasting Formula $S' = \sqrt{\left\{\frac{\Sigma\left(Y/\bar{Y}_5 - C/\bar{C}_5\right)^2}{N}\right\}}$				
		May	June	July	August	September
Texas	24.64	26.38	22.11	19.23	17.86	13.77
Georgia	13.89	17.28	15.20	12.17	11.92	11.08
Alabama	13.37	17.59	13.58	12.24	11.66	10.21
South Carolina	18.65	21.90	19.06	17.02	19.03	15.28

If, now, we examine the results that are collected in Table 13, we see that the May report as to the condition of the crop, which is issued about the first of June, not only has no value in case of all four of the representative states, but that it is worse than useless since in all four cases the value of S' is greater than σ_y. The forecasts by means of the official formula miss the actual yield-ratios by an amount S' which exceeds σ_y, the value of the variability of the actual yield-ratios when no forecast is made at all. The June reports, which are issued about the first of July, are in three out of four cases worse than useless, because in case of three of the states S' is greater than σ_y. The reports for July, August, and September have real value as forecasts and the value increases as the crop approaches maturity.

In the preceding chapter it was made clear that an improvement upon the official results would be obtained if, as a forecasting formula, we should take the equation $(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$, where y is the value of the predicted yield-ratio $Y/\bar{Y}_5$; $\bar{y}$ is the mean value of the actual yield-ratios $Y/\bar{Y}_5$ for the whole period under investigation; r is the coefficient of correlation; σ_y and σ_x are, respectively, the variabilities of $Y/\bar{Y}_5$ and $C/\bar{C}_5$; x is the value of $C/\bar{C}_5$ for the current year; and $\bar{x}$ is the mean value of $C/\bar{C}_5$ for the whole period under investigation. We know from the theory of Chapter II, "The Mathematics of Correlation," that when the above equation is used as a prediction formula, the degree of accuracy of the prediction is measured by $S = \sigma_y\sqrt{1 - r^2}$; that is to say, the value of S gives the root-mean-square of the deviations of the actual yield-ratios from the predicted yield-ratios.

The relative accuracy of this formula as compared with the official formula is exhibited by the calculations in Table 14.

In Table 14, r measures the degree of correlation between the forecast series $C/\bar{C}_5$ and the actual yield-ratios $Y/\bar{Y}_5$; S, as we have indicated, measures the scatter of the actual yield-ratios about the predicted yield-ratios, when the forecast is made by means of the formula $(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$; S' measures the scatter of the actual yield-ratios about the predicted yield-ratios, when the forecasts are made by means of the official formula.

Table 14 shows, by the magnitude of r, that the May reports have no value and that the reports for the other

TABLE 14. — Relative Accuracy of Forecasting Formulas

(1) The Correlation Equation $(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$; the Official Formula $C/\bar{C}_5 = Y/\bar{Y}_5$

The Representative States	May			June			July			August			September		
	r	S	S'	r	S	S'	r	S	S'	r	S	S'	r	S	S'
Texas	.046	24.61	26.38	.449	22.02	22.11	.684	17.98	19.23	.710	17.35	17.86	.838	13.45	13.77
Georgia	.047	13.87	17.28	.325	13.14	15.20	.660	10.43	12.17	.664	10.39	11.92	.743	9.30	11.08
Alabama	— .037	13.36	17.59	.438	12.02	13.58	.576	10.93	12.24	.625	10.44	11.66	.726	9.19	10.21
South Carolina	.038	18.64	21.90	.364	17.38	19.06	.568	15.35	17.02	.328	17.62	19.03	.688	13.53	15.28

months have increasing value as the crop approaches maturity. The comparative values of S and S' show that in every month, in all four of the representative states, the correlation equation gives a more accurate forecast than the official formula. Moreover, the correlation equation as a forecasting formula does not admit of the anomalous results which were brought out in the consideration of Table 13, where we found that in a number of cases S' was greater than σ_y, which signifies that the forecasts are worse than useless. When the correlation equation is employed as a forecasting formula it is impossible for S to exceed σ_y, since $S = \sigma_y\sqrt{1 - r^2}$.

The results collected in the two Tables in this section establish two of the theses enunciated at the beginning of the chapter:

(1) That some of the official reports referring to the representative states are valuable as forecasts, but that others are worse than useless in the sense of supplying erroneous instruction as to the crop outlook, and thereby suggesting a misdirecting of activity on the part of farmers, dealers, and manufacturers;

(2) That even in case of the useful forecasts the official method does not extract the full amount of truth contained in the laboriously collected data.

Forecasting the Yield of Cotton from the Accumulated Effects of the Weather [1]

Throughout the period from the first of May until the end of September, the growth of the cotton plant is

[1] Professor J. Warren Smith, of the University of Ohio, and Mr. R. H. Hooker, of London, have been pioneers in dealing with special phases

watched with anxious solicitude. The changes of the weather may convert a crop that is flourishing at the first of June into a comparative failure at the time of harvest, or the damage of excessive heat in July may be off-set by a beneficial rainfall in August. The effects of temperature and rainfall upon the crop vary from state to state, but in all cases the effects are cumulative, and the probable consequences of a rain or drought at any point in the growth season are dependent upon the quantity and distribution of rainfall and temperature preceding the time in question. The principal difficulty in forecasting the yield of the crop from the changes in the weather is that there are so many variables in the problem and all of the variables are interrelated. In a particular state it may be that the growing plant needs a cool, dry July and a rainy, hot August; but temperature and rainfall may be so interrelated that, on the average, in July when the weather is cool, it is likewise rainy, and in August when the weather is hot, it is also dry; and it might be more important that the crop should have rain in August than that July should be cool. The economist who, at any given time in the growth period, seeks to forecast the yield of cotton at harvest must be able to measure the effects upon the crop of the accumulated variations in temperature and rainfall up to the time in question.

Before passing to the account of the method adopted to measure the accumulated effect of the weather upon

of the topic treated in this section. Professor Smith was one of the first to see the economic importance of forecasting the yield of the crops from the weather, and Mr. Hooker, as far as I know, led the way in the use of the method of multiple correlation to measure the joint effect of temperature and rainfall upon the yield of the crops.

the growing crop, we shall consider the device for meeting the difficulties that are traceable to the secular trend and cyclical variations in the yield per acre, the rainfall, and the temperature. In some states the yield per acre of cotton throughout the period under investigation has steadily increased, while in other states it has steadily decreased. Moreover, in all of the states the yield per acre, temperature, and rainfall have, during the same interval, been subjected to cyclical influences. In order to measure the relation between the yield per acre and the variations in the weather, we must make allowance for these secular and cyclical changes, and to do this we have profited by the experience of the Bureau of Statistics of the Department of Agriculture. We recall that, in order to forecast from the condition of the crop in any month the probable yield of the crop at the end of the year, the Bureau of Statistics does not work directly with the absolute values of the condition and the yield, but it takes the condition-ratio and the yield-ratio, the forecasting formula being $C/\bar{C}_5 = Y/\bar{Y}_5$. In the preceding chapter we showed that equally good results would be obtained by using the formula $C/\bar{C}_3 = Y/\bar{Y}_3$; that is to say, instead of making the denominators five years averages, to employ three years averages. One advantage of the latter method is that when the available data are few and as many as possible must be utilized, the three years method gives a larger number of cases upon which to base one's computations.

We shall use this three years method in correlating the temperature-ratios and rainfall-ratios of the several months with the yield-ratios of cotton. For each month

the series to be correlated will be $T/\bar{T}_3, Y/\bar{Y}_3; R/\bar{R}_3, Y/\bar{Y}_3$. In these formulæ T is the average temperature for the given month, and $\bar{T}_3$ is the average temperature for the same month during the preceding three years; R is the amount of rainfall for the given month, and $\bar{R}_3$ is the average amount of rainfall for the same month during the preceding three years; Y is the yield per acre of cotton for the given year, and $\bar{Y}_3$ is the average yield per acre of cotton for the three years preceding the given year. In Table 15 the method of preparing the data for computing the correlation is illustrated by the correlation, in Georgia, between the June temperature-ratio and the yield-ratio of cotton.

When the two series that are given in columns VIII and IX of Table 15 are correlated, it is found that $r = .551$, and this value of r gives for the value of the scatter, $S = \sigma_y\sqrt{1 - r^2} = 13.89\sqrt{1 - r^2} = 11.59$. By referring to Table 13 we find that the official method of forecasting from the condition of the crop gives for the month of June, in Georgia, $S' = 15.20$. The official forecast, inasmuch as $\sigma_y = 13.89$, is worse than useless, while the forecast of the crop from the June temperature has a decided value. We also know from Table 13 that the official method of forecasting from the condition of the crop gives for the month of May, in Georgia, a value of $S' = 17.28$, which is also worse than useless. But the correlation between the May rainfall-ratio in Georgia and the yield-ratio in Georgia gives $r = -.410$, and $S = 12.67$. These two illustrations show that the cotton crop in Georgia is favorably affected by a dry May and a warm June. How would it be possible to utilize the knowledge both of the rainfall in May and

TABLE 15. — GEORGIA. CORRELATION BETWEEN THE JUNE TEMPERATURE-RATIO AND THE YIELD-RATIO OF COTTON

I Year	II June Mean Temperature T	III Sum for Preceding Three Years in Column II	IV Mean Temperature for Three Years Column III Divided by Three $\overline{T}_3$	V Yield per Acre of Cotton in Pounds of Lint Y	VI Sum for Preceding Three Years in Column V	VII Mean Yield per Acre of Cotton Column VI Divided by Three $\overline{Y}_3$	VIII June Temperature Ratio $T/\overline{T}_3$	IX Yield Ratio $Y/\overline{Y}_3$
1892	78.1			160				
3	74.9			136				
4	77.1			155				
5	77.9	230.1	76.7	152	451	150	101.6	101.3
6	77.7	229.9	76.6	122	443	148	101.4	82.4
7	80.8	232.7	77.6	178	429	143	104.1	124.5
8	80.1	236.4	78.8	183	452	151	101.6	121.2
9	80.2	238.6	79.5	159	483	161	100.9	98.8
1900	75.6	241.1	80.4	172	520	173	94.0	99.4
1	77.6	235.9	78.6	167	514	171	98.7	97.7
2	79.5	233.4	77.8	165	498	166	102.2	99.4
3	74.4	232.7	77.6	158	504	168	95.9	94.0
4	77.4	231.5	77.2	205	490	163	100.3	125.8
5	78.7	231.3	77.1	200	528	176	102.1	113.6
6	77.8	230.5	76.8	165	563	188	101.3	87.8
7	76.0	233.9	78.0	190	570	190	97.4	100.0
8	77.5	232.5	77.5	190	555	185	100.0	102.7
9	78.5	231.3	77.1	184	545	182	101.8	101.1
10	75.3	232.0	77.3	173	564	188	97.4	92.0
11	80.9	231.3	77.1	240	547	182	104.9	131.9
12	75.4	234.7	78.2	163	597	199	96.4	81.9
13	76.4	231.6	77.2	208	576	192	99.0	108.3
14	82.2	232.7	77.6	239	611	204	105.9	117.2

the temperature in June to forecast, at the end of June, the probable yield of the crop? This question brings us to a consideration of the method of multiple correlation.

In Chapter II, "The Mathematics of Correlation," we developed in considerable detail the theory of the correlation between two variables. The essential steps are:

(1) The assumption[1] that the two variables are related in a linear way by the equation $y = mx + b$;

(2) The calculation of the coefficient of correlation r; and

(3) The determination of the accuracy of the equation $y = mx + b$ as a forecasting formula by calculating the scatter $S = \sigma_y \sqrt{1 - r^2}$. In the theory of multiple correlation the essential steps run parallel to those in the theory of the correlation of two variables. In case of three variables the three steps are:

(1) The assumption that the equation connecting the three variables,

$$x_0,\ x_1,\ x_2 \text{ is } x_0 = a_0 + a_1 x_1 + a_2 x_2;$$

(2) The determination of the degree of association between the variable x_0 and the other two variables x_1, x_2 by calculating the coefficient of multiple correlation R;

(3) The determination of the accuracy of the equation $x_0 = a_0 + a_1 x_1 + a_2 x_2$ as a forecasting formula by calculating the scatter, $S'' = \sigma_0 \sqrt{1 - R^2}$.

We found, in the theory of the correlation of two vari-

[1] There are methods for testing the legitimacy of the assumption which should, of course, be applied.

ables, that the forecasting formula, namely, $y = mx + b$, may be put into the form $(y - \bar{y}) = m(x - \bar{x})$. In a similar manner, when three variables x_0, x_1, x_2, are correlated, the forecasting formula may be put into the form

$$(x_0 - \bar{x}_0) = a_1(x_1 - \bar{x}_1) + a_2(x_2 - \bar{x}_2).$$

Our statistical problem will be solved if we can determine from the statistical data the following three items:

(1) The values of the coefficients of $(x_1 - \bar{x}_1)$ and $(x_2 - \bar{x}_2)$ in the forecasting formula.

These values are

$$a_1 = \frac{r_{01} - r_{02}r_{12}}{1 - r_{12}^2}\frac{\sigma_0}{\sigma_1}; \quad a_2 = \frac{r_{02} - r_{01}r_{12}}{1 - r_{12}^2}\frac{\sigma_0}{\sigma_2}.$$

Here r_{01} is the coefficient of correlation between the variables x_0, x_1; r_{02} is the coefficient of correlation between x_0 and x_2; r_{12} the correlation between x_1 and x_2; σ_0 is the standard deviation of the x_0's; σ_1 is the standard deviation of the x_1's; and σ_2 the standard deviation of the x_2's. When these values of a_1, a_2 are substituted in the forecasting formula $(x_0 - \bar{x}_0) = a_1(x_1 - \bar{x}_1) + a_2(x_2 - \bar{x}_2)$, the most probable value of x_0 may be calculated from the known values of x_1, x_2.

(2) The value of the coefficient of multiple correlation, R. In case of three variables x_0, x_1, x_2,

$$R^2 = \frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2}.$$

(3) The value of the scatter, $S'' = \sigma_0\sqrt{1 - R^2}$, which measures the root-mean-square of the deviations of the observed values of x_0 from the most prob-

able values of x_0 when the most probable values are predicted from the forecasting formula $(x_0 - \bar{x}_0) = a_1(x_1 - \bar{x}_1) + a_2(x_2 - \bar{x}_2)$.

We may proceed at once to illustrate the method by showing how the yield-ratio of cotton, in Georgia, may be predicted from the rainfall-ratio for May and the temperature-ratio for June. Let the yield-ratio be x_0, the rainfall-ratio for May be x_1, and the temperature-ratio for June be x_2. The forecasting formula is

$$(x_0 - \bar{x}_0) = a_1(x_1 - \bar{x}_1) + a_2(x_2 - \bar{x}_2).$$

From the statistical data we find that

(1) The mean values of x_0, x_1, x_2, are, respectively,

$$\bar{x}_0 = 104.05;\ \bar{x}_1 = 107.04;\ \bar{x}_2 = 100.34;$$

(2) The standard deviations of x_0, x_1, x_2 are, respectively,

$$\sigma_0 = 13.890;\ \sigma_1 = 65.528;\ \sigma_2 = 3.142;$$

(3) The coefficient of correlation between x_0 and $x_1 = r_{01} = -.410$; between x_0 and $x_2 = r_{02} = .551$; between x_1 and $x_2 = r_{12} = -.427$.

Since $a_1 = \dfrac{r_{01} - r_{02}r_{12}}{1 - r_{12}^2}\dfrac{\sigma_0}{\sigma_1}$, and $a_2 = \dfrac{r_{02} - r_{01}r_{12}}{1 - r_{12}^2}\dfrac{\sigma_0}{\sigma_2}$, the substitution of the above numerical values for the algebraic symbols gives $a_1 = -.045$, and $a_2 = 2.033$. After the proper substitutions have been made and the equation simplified, the forecasting formula becomes $x_0 = -95.12 - .045x_1 + 2.033x_2$.

Since $R^2 = \dfrac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2}$, we get for the coefficient of correlation between x_0 and the two variables x_1, x_2,

the value $R^2 = .340933$, or $R = .584$. Furthermore, since $S'' = \sigma_0\sqrt{1 - R^2}$, we get as the numerical measure of the accuracy of the forecasting formula, $S'' = 11.28$.

We have seen that, by means of the forecast from the weather, we get a formula for reducing the variability of σ_0 even in May, while the Government report for May we have found to be erroneous and misleading. Furthermore, by means of the additional information given by the weather reports for June, we have been able still further to reduce the variability of σ_0. The value of $S'' = \sigma_0\sqrt{1 - R^2}$ for June is found to be 11.28. We may at this point take stock of our gains. From Table 13 we know that, according to the official method, the accuracy of the forecasts are, for May, $S' = 17.28$; for June, $S' = 15.20$; for July, $S' = 12.17$; for August, $S' = 11.92$; and for September, $S' = 11.08$. But by means of the method of forecasting from the data of the weather, we get, for May, $S'' = 12.67$; and for June, $S'' = 11.28$, where, in the June forecast, we use the rainfall-ratio for May and the temperature-ratio for June. Not only are the forecasts from the weather for these two months better than the forecasts by the official method from the condition of the crop, but the value of S'' for May is about as good as the value of S' two months later, at the end of July; and the value of S'' for June is about as good as the value of S' two or three months later at the end, respectively, of August and September.

But our method admits of still further usefulness. From the accompanying Table 16, we see that the fruitfulness of the cotton crop is affected not only

by the weather of May and June but also by that of July and of August. The coefficients for both temperature and rainfall in August have significant values. If at the end of August we should wish to forecast the yield of cotton from accumulated effects of past weather — for example, of the May rainfall, the June temperature, and the August temperature — we have only to extend the principles of multiple correlation to cover four variables. The steps in the development are again three in number and they run parallel with the three steps that we have already traversed in describing the correlation between two variables and the correlation between three variables.

TABLE 16. — GEORGIA. CORRELATION BETWEEN THE YIELD-RATIO OF COTTON AND THE TEMPERATURE-RATIO AND RAINFALL-RATIO

	Values of the Coefficient of Correlation				
	May	June	July	August	September
Temperature-Ratio	—.097	.551	—.032	—.499	.082
Rainfall-Ratio	—.410	—.411	—.254	.426	—.188

The three steps are:

(1) The assumption that the equation connecting the four variables x_0, x_1, x_2, x_3 is $x_0 = a_0 + a_1x_1 + a_2x_2 + a_3x_3$. It is not difficult to prove that this forecasting equation may be put into the form

$$(x_0 - \bar{x}_0) = a_1(x_1 - \bar{x}_1) + a_2(x_2 - \bar{x}_2) + a_3(x_3 - \bar{x}_3).$$

By the *Method of Least Squares* the values of the coefficients in the forecasting equation are ascertained to be

$$a_1 = \frac{r_{01}(1 - r_{23}^2) + r_{02}(r_{13}r_{23} - r_{12}) + r_{03}(r_{12}r_{23} - r_{13})}{(1 - r_{23}^2) + r_{12}(r_{13}r_{23} - r_{12}) + r_{13}(r_{12}r_{23} - r_{13})} \frac{\sigma_0}{\sigma_1};$$

$$a_2 = \frac{r_{02}(1 - r_{13}^2) + r_{03}(r_{12}r_{13} - r_{23}) + r_{01}(r_{13}r_{23} - r_{12})}{(1 - r_{13}^2) + r_{23}(r_{12}r_{13} - r_{23}) + r_{12}(r_{13}r_{23} - r_{12})} \frac{\sigma_0}{\sigma_2};$$

$$a_3 = \frac{r_{03}(1 - r_{12}^2) + r_{02}(r_{12}r_{13} - r_{23}) + r_{01}(r_{12}r_{23} - r_{13})}{(1 - r_{12}^2) + r_{23}(r_{12}r_{13} - r_{23}) + r_{13}(r_{12}r_{23} - r_{13})} \frac{\sigma_0}{\sigma_3}.$$

(2) The determination of the degree of association between the variable x_0 and the other three variables by calculating the coefficient of multiple correlation R. In the case of four variables the value of R is given by the following equation:

$$R^2 = \frac{M}{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}, \text{ where}$$

$$M = \left\{ \begin{array}{c} (r_{01}^2 + r_{02}^2 + r_{03}^2 - r_{01}^2r_{23}^2 - r_{02}^2r_{13}^2 - r_{03}^2r_{12}^2) \\ -2(r_{01}r_{02}r_{12} + r_{01}r_{03}r_{13} + r_{02}r_{03}r_{23}) \\ +2(r_{01}r_{02}r_{13}r_{23} + r_{01}r_{03}r_{12}r_{23} + r_{02}r_{03}r_{12}r_{13}) \end{array} \right\}$$

(3) The determination of the accuracy of $(x - \bar{x}_0) = a_1(x_1 - \bar{x}_1) + a_2(x_2 - \bar{x}_2) + a_3(x_3 - \bar{x}_3)$ as a forecasting formula by calculating the scatter, $S'' = \sigma_0\sqrt{1 - R^2}$.

To illustrate the use of this more complex forecasting formula we shall go through the work of ascertaining how the yield-ratio x_0 may be predicted from the knowledge of three variables, x_1 = May rainfall-ratio, x_2 = June temperature-ratio, x_3 = August temperature-ratio.

From the statistical data we find that

(1) The mean values of x_0, x_1, x_2, x_3 are, respectively,

$\bar{x}_0 = 104.05$; $\bar{x}_1 = 107.04$; $\bar{x}_2 = 100.34$; $\bar{x}_3 = 100.15$;

(2) The standard deviations of x_0, x_1, x_2, x_3 are, respectively,

$\sigma_0 = 13.890$; $\sigma_1 = 65.528$; $\sigma_2 = 3.142$; $\sigma_3 = 1.761$;

(3) The coefficients of correlation are

$$r_{01} = -.410;\ r_{02} = .551;\ r_{03} = -.499;\ r_{12} = -.427;$$
$$r_{13} = .014;\ r_{23} = -.126.$$

When these numerical values are substituted for the algebraic symbols in the above formulæ, we obtain for the forecasting formula, $x_0 = 286.84 - .050x_1 + 1.743x_2 - 3.518x_3$; for the coefficient of multiple correlation, $R = .732$; for the scatter, $S'' = 9.46$.

In Figure 9 the continuous line represents the values of the yield-ratios for the several years as the yield-ratios are computed from the actual statistics by means of the formula $Y/\bar{Y}_3$; the dashed line represents the yield-ratios as they are computed from the rainfall-ratio of May, the temperature-ratio of June, and the temperature-ratio of August by means of the forecasting formula

$$x_0 = 286.84 - .050x_1 + 1.743x_2 - 3.518x_3;$$

The root-mean-square deviation of the actual ratios from the predicted ratios is $S'' = 9.46$.

Figure 10 illustrates the degree of precision in the forecast at the end of August, by means of the official method, of the yield of cotton in Georgia. The continuous line represents the value of the actual yield-ratios, computed by the formula $Y/\bar{Y}_5$; and the dashed line represents the theoretical yield-ratios as they are predicted by the formula $C/\bar{C}_5$. The root-mean-square deviation of the actual ratios from the predicted ratios is $S' = 11.92$.

If now we compare the value of S'' for August with the value of S' for September, that is, $S' = 11.08$,

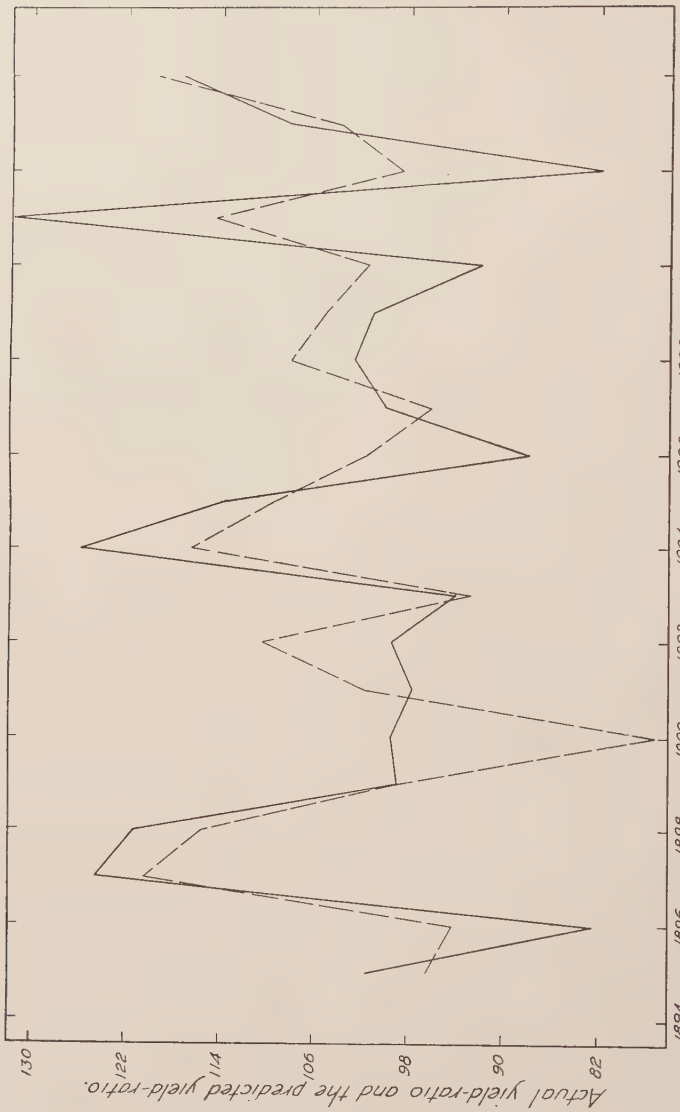

FIGURE 9. — Georgia. The actual yield-ratios of cotton, —, and the yield-ratios as predicted, at the end of August, from the accumulated effects of the weather, - - -.

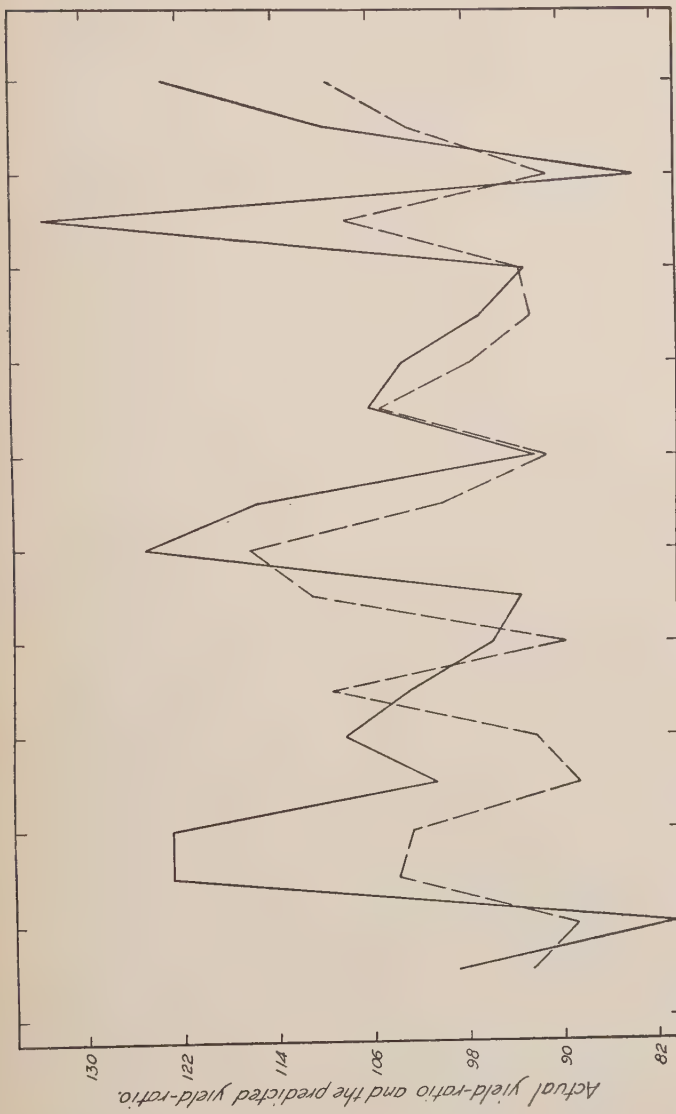

FIGURE 10. — Georgia. The actual yield-ratios of cotton, —, and the yield-ratios as predicted, by the official formula, from the condition of the crop at the end of August, – – –.

we see that from the weather data we can obtain, by means of mathematical methods, a better forecast at the end of August than the official method enables us to obtain from the condition of the crop at the end of September. A clearer view of the value of forecasting the cotton yield from the weather reports, by means of the methods that we have described, may be had from the following Table 17.

TABLE 17. — Relative Accuracy of the Forecasts of the Yield Per Acre of Cotton (1) from the Condition of the Crop, by the Official Method, and (2) from the Weather Reports, by the Method of Correlation. Georgia

		Months				
		May	June	July	August	September
Error of the Forecasts	S'	17.28	15.20	12.17	11.92	11.08
	S''	12.67	11.28	9.94	9.46	9.46

In computing S'' we used for May, the rainfall-ratio; for June, the May rainfall-ratio and the June temperature-ratio; for July, the May rainfall-ratio, the June temperature-ratio, and the July rainfall-ratio; for August and September, the May rainfall-ratio, the June temperature ratio, and the August temperature-ratio. From Table 17 it is seen that not only is S'' less than S' for every month, but S'' for May is about as good as S' two months later, at the end of July; the S'' for June is about as good as the S' two months later, at the end of August; the S'' for July is better than the S' two months later, at the end of September; and the S'' at the end of August is better than the S' at the end of September.

As far as the state of Georgia is concerned we have proved our thesis: That it is possible, by means of the weather reports and mathematical methods, to forecast the yield per acre of cotton with a greater degree of precision than the reports of the official Bureau with its vast organization for the collection and reduction of data referring to the condition of the growing crop.

The Results Compared for the Representative States

The thesis that has just been proved for the state of Georgia we shall now test for the representative states Texas, Georgia, Alabama, and South Carolina, which together, in 1914, produced 65 per cent of the total crop of the United States. In Table 18 are exhibited the correlations between the yield-ratios of cotton and the temperature-ratios and rainfall-ratios of the representative states.[1] If we refer to the earlier part

[1] The statistics of the condition of the crop and of the yield per acre of cotton were kindly supplied to me by Mr. Leon M. Estabrook and Mr. George K. Holmes of the U. S. Department of Agriculture. The figures are reproduced in Tables 23, 24, 25, 26 of the Appendix to this chapter. The weather data for Georgia, Alabama, and South Carolina were taken from the publication of the U. S. Weather Bureau *Climatological Data* for 1915, and refer, in each case, to the mean temperature and mean rainfall for the entire state. The coefficients of correlation between the yield-ratios and the weather-ratios were based upon 20 ratios in case of Georgia (1895–1914); 17 ratios in case of Alabama (1898-1915); 21 ratios in case of South Carolina, and 21 ratios in case of Texas (1894–1914). Because of the great size of Texas and the concentration of the cotton production in the Eastern and Central parts of the state, the mean temperature and mean rainfall were computed for those two sections from the records for the individual stations that are given in the *Annual Reports* of the Chief of the U. S. Weather Bureau. The thirty-one stations that were selected for the rainfall record were: Abilene, Albany, Austin, Brenham, Brownwood, Claytonville, Coleman, College Station, Corsicana, Dallas, Fairland, Fort Worth, Fredericksburg, Gainesville, Graham, Greenville, Huntsville,

TABLE 18. — The Correlation Between the Yield-Ratio of Cotton and the Temperature-Ratio and the Rainfall-Ratio in Representative States

Representative States	Values of the Coefficient of Correlation									
	May		June		July		August		September	
	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall
Texas	.163	.056	—.047	.195	—.474	—.057	—.628	.691	.133	—.078
Georgia	—.097	—.410	.551	—.411	—.032	—.254	—.499	.426	.082	—.188
Alabama	.185	—.568	.247	—.385	—.165	.258	—.277	.222	.095	—.242
South Carolina	.130	—.485	.409	—.361	—.098	.116	—.381	.370	.065	.057

of this chapter we shall see that the test of the accuracy of the official crop reports rests upon the yield-ratios and condition-ratios for the period 1894–1914. In order that the accuracy of the forecasts from the weather may be more fairly compared with the accuracy of the forecasts from the condition of the crop, the period of the weather observations has been taken, in case of each state, as near as possible to the period 1894–1914. The Texas ratios are from 1894 to 1914; those of Georgia from 1895 to 1914; of South Carolina, 1894 to 1914. Because of the limited meteorological record in Alabama the longest series of ratios that could be obtained runs from 1898 to 1915.

The raw data are given in Tables 27, 28, 29, 30 of the Appendix to this chapter, and in computing the coefficients of correlation all of the data in the Tables have been used exactly as they are recorded.[1] It would have been possible, on several occasions, to increase the coefficients by omitting one or two rainfall-ratios which, in consequence of torrential storms, presented unduly large values; but no such liberty has been taken with the crude material, although for purposes of forecasting the yield in normal times such a procedure might have been justifiable.

Lampasas, Longview, Menardville, Nacogdoches, Palestine, Panter, Paris, San Angelo, Sulphur Springs, Taylor, Temple, Tyler, Waco, Weatherford. The seventeen stations the records of which were used in compiling the mean temperature were Abilene, Brenham, Brownwood, Corsicana, Dallas, Fort Worth, Greenville, Huntsville, Lampasas, Longview, Nacogdoches, Palestine, Paris, Taylor, Temple, Waco, Weatherford.

The weather data for all four states are given in Tables 27, 28, 29, 30, of the Appendix to this chapter.

[1] The omission of the August rainfall for 1914, in Texas, is recorded in the Notes to Table 19.

A summary view of the relative accuracy of the forecast of the yield from the condition of the crop and the forecast from the accumulated rainfall and temperature is given in Table 19. In considering the following comments on Table 19 we shall bear in mind that S' measures the accuracy of the forecasts from the condition of the crop by the official method, and S'', the accuracy of the forecast from the weather by the method of correlation. The more accurate the forecasts, the smaller are the respective values of S' and S''.

(1) There are four representative states — Texas, Georgia, Alabama, and South Carolina, and there are five monthly reports on the condition of the growing crop, the reports describing the condition of the crop at the end of May, June, July, August, and September. There are, therefore, twenty cases in which the accuracy of the two methods may be compared. Table 19 shows that in 17 out of 20 cases the forecast from the weather by the method of correlation is more accurate than the forecast by the official method from the condition of the growing crop.

(2) For all of the representative states the forecasts by the official method from the May condition of the crop are worse than useless because the values of S' are larger than the corresponding values of σ_y. On the contrary, the forecasts from the May weather by the method of correlation have a real value.[1] The forecasts from the weather for Georgia and South Carolina are, at the end of May, better than the official forecasts for June, and nearly as good as the official forecasts at the end of July; and the forecast for Alabama at the

[1] The last of the Notes on Table 19 should be consulted.

end of May is nearly as good as the official forecast at the end of September.

TABLE 19. — Relative Accuracy of the Forecasts of the Yield Per Acre of Cotton (1) from the Condition of the Crop, by the Official Method, and (2) from the Weather Reports, by the Method of Correlation

Representative States	Standard Deviation of the Yield-Ratios	Error of the Forecast from the Condition of the Crop, S' Error of the Forecast from the Weather, S''									
		May		June		July		August		September	
	σ_y	S'	S''	S'	S''	S'	S''	S'	S''	S'	S''
Texas	24.64	26.38	25.31	22.11	25.15	19.23	22.58	17.86	16.80	13.77	16.80
Georgia	13.89	17.28	12.67	15.20	11.28	12.17	9.94	11.92	9.46	11.08	9.46
Alabama	13.37	17.59	10.40	13.58	9.70	12.24	9.52	11.66	9.19	10.21	9.19
South Carolina	18.65	21.90	17.42	19.06	16.10	17.02	16.10	19.03	14.98	15.28	14.98

In computing S', the formula $S' = \sqrt{\frac{\Sigma\left(Y/\bar{Y}_3 - C/\bar{C}_3\right)^2}{N}}$ was used for every state and every month.

In obtaining S'' the formulas $S'' = \sigma_0\sqrt{1-r^2}$, $S'' = \sigma_0\sqrt{1-R^2}$ were used according as one or more independent variables were employed. The combinations of variables were:

In case of Texas: For May, temperature-ratio; for June, rainfall-ratio; for July, temperature-ratio; for August and September, July temperature-ratio, August temperature ratio, and August rainfall ratio. In computing the correlation between the yield-ratio and the rainfall-ratio for August, the rainfall data for 1914 were not used.

In case of Georgia: For May, rainfall-ratio; for June, May rainfall-ratio and June temperature-ratio; for July, May rainfall-ratio, June temperature-ratio, and July rainfall-ratio; for August and September, May rainfall-ratio, June temperature-ratio, August temperature-ratio.

In case of Alabama: For May, rainfall-ratio; for June, May rainfall-ratio, June rainfall-ratio; for July, May rainfall-ratio, June rainfall-ratio, July rainfall-ratio; for August and September, May rainfall-ratio, June rainfall-ratio, August temperature-ratio.

In case of South Carolina: For May, rainfall-ratio; for June and July, May rainfall-ratio, June rainfall-ratio and June temperature-ratio; for August and September, May rainfall-ratio, June temperature-ratio, August temperature-ratio.

The ratios that were correlated were obtained from the official statistics by means of the formulas $T/\bar{T}_3$, $R/\bar{R}_3$, $Y/\bar{Y}_3$, where the symbols refer, respectively, to the temperature, rainfall, and yield per acre of cotton.

The values of σ_y are the standard deviations of the yield-ratios when the five years progressive means are used. This will explain the rather anomalous result that S'', in case of Texas, is for May and June larger than σ_y. When the three years means are the basis of the yield-ratios, the standard deviation is 25.65.

(3) For three out of the four states the prediction from the June condition of the crop by the official method is worse than useless since the values of S' are greater than the corresponding values of σ_y. But

the forecasts from the weather are in all three cases of decided value, being in all three cases better than the official forecasts for the following month.

(4) For all of the states except Texas the forecast from the weather gives, for each month, a more accurate prediction than can be obtained by the official method from the condition of the crop *one month later*. That is to say, for all of the states except Texas, S'' for May is smaller than S' for June; S'' for June is smaller than S' for July; S'' for July is smaller than S' for August; and S'' for August is smaller than S' for September.

(5) For all of the states except Texas the forecasts from the weather give for May, June, and July about as good predictions as can be obtained by the official method from the condition of the crop *two months later*. (In six out of the nine possible cases S'' is less than S' *two months later*.)

Considering the character of these findings it is clear that, as far as concerns the representative states which produce sixty-five per cent of the total cotton crop of the United States, we may conclude, in terms of our thesis: "Notwithstanding the vast official organization for collecting and reducing data bearing upon the condition of the growing crop, it is possible, by means of mathematical methods, to make more accurate forecasts than the official reports, in the matter of the prospective yield per acre of cotton, simply from the data supplied by the Weather Bureau as to the current records of rainfall and temperature in the respective cotton states."

Three Possible Objections

The substance of this section, which is technical in character, is intended to meet three quite natural objections:

(1) In the preceding chapter and in the early part of the present chapter, the defects in the official forecasting formula were pointed out, and the method of correlation, with the equation $(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$, was shown to give better results in all of the representative states and in every month of the growing season. If this better forecasting formula were applied to the official data referring to the condition of the growing crop, would not the forecasts of yield per acre be better than the forecasts which we have been able to make from the current records of temperature and rainfall in the cotton states?

(2) Although data as to the condition of the growing cotton crop have been officially collected and published since 1866, the official Bureaus refrained, until 1911, from interpreting their own data in the definite form of quantitative forecasts. May not the defects in the official forecasts which we have located and measured be due to the fact that the official prediction formula, which was promulgated in 1911, has been applied in this Essay to data running through a quarter of a century?

(3) The problem of measuring the relation between the yield per acre of cotton, and the amount of rainfall and the temperature at various epochs, in its period of growth, presents theoretical and practical difficulties

that leave any attempt at solution open to possible objections. The chief difficulties are (1) that the series of available data — the yield per acre series and the series of rainfall and temperature records — are summary results of three classes of changes with three different sets of causes: (a) secular changes, (b) cyclical changes, (c) random changes; and (2) that there is no known statistical method that will enable one with series as short as ours to segregate satisfactorily the effects of these three types of changes. May it not be true, therefore, that the good results which we have obtained in our forecasts from the weather are results that do not rest upon real causes but are largely spurious, inhering in the method which we have employed?

We shall proceed to the consideration of these three objections. Table 20 presents the data that are necessary to compare the accuracy of the forecasts from the official material as to the condition of the crop, and the forecasts from the records of temperature and rainfall, both of the forecasts being made by the methods of correlation. S measures the accuracy of the forecast from the condition of the crop, where $S = \sigma_y \sqrt{1 - r^2}$, and the forecasting equation is $(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$. S'' measures the accuracy of the forecast from the weather and has the same value as it has retained throughout the investigations of this chapter. The smaller the values of S, S'', the more accurate are the respective forecasts. An examination of Table 20 shows that since there are four representative states and five months of the growing season, there are 20 cases in which the values of S and S'' may be com-

pared. In 12 out of these 20 cases S'' is smaller than S. For purposes of exploiting the prospects of the crop, the earlier a reliable forecast can be obtained, the greater is its economic value. If, therefore, we omit the month of September, there are, in Table 20, 16 cases in which S and S'' may be compared, and in 12 of these 16 cases S'' is smaller than S.

TABLE 20. — Relative Accuracy of the Forecasts of the Yield Per Acre of Cotton (1) from Data as to the Condition of the Crop, by Means of the Correlation Equation; (2) from Data as to the Weather, by Means of the Method of Multiple Correlation

Representative States	Error of the Forecast from the Condition of the Crop, S Error of the Forecast from the Weather, S''									
	May		June		July		August		September	
	S	S''	S	S''	S	S''	S	S''	S	S''
Texas	24.61	25.31	22.02	25.15	17.98	22.58	17.35	16.80	13.45	16.80
Georgia	13.87	12.67	13.14	11.28	10.43	9.94	10.39	9.46	9.30	9.46
Alabama	13.36	10.40	12.02	9.70	10.93	9.52	10.44	9.19	9.19	9.19
South Carolina	18.64	17.42	17.38	16.10	15.35	16.10	17.62	14.98	13.53	14.98

We conclude from these comparisons that, as far as concerns the representative states producing sixty-five per cent of the total cotton crop, the forecasts from the weather are at least as good as the forecasts from the condition of the crop by means of the formula $(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$. This latter formula, we have already seen, gives a better forecast than the official formula in all of the months of the growing season of cotton, in all of the representative states.

Table 21 contains the material by means of which an opinion may be formed as to the proper answer to the

second of our three questions. Throughout this Essay the error of the official formula has been measured by the scatter of the forecasts, $S' = \sqrt{\left\{\frac{\Sigma(Y/\bar{Y}_5 - C/\bar{C}_5)^2}{N}\right\}}$, and in case of each of the representative states, the computed value of S' rested upon the observations for the 21 years, 1894 to 1914. In 1911 the Department of Agriculture defined its formula for predicting, from the monthly condition of the crop, the ultimate yield per acre of cotton; but until that year the condition figures were published without an official attempt to suggest what definite inference, as to the ultimate yield, should be made from the crop reports. It might therefore be assumed that the year 1911 would mark the beginning of a more accurate crop-reporting service, and that, under the new procedure of the Department of Agriculture, the conclusions which we have drawn concerning the comparative accuracy of forecasts from the condition of the crop by means of the official formula, and forecasts from the weather by means of correlation equations, would no longer hold true. Or one might urge that throughout the entire period 1894–1914 the crop-reporting service had been continuously improved, and that, consequently, the precision of the official formula when measured by the record for this long period would be no index of its present accuracy. These very natural doubts raise two important questions of fact: (a) Has the continuous improvement of the crop-reporting service throughout the 21 years 1894–1914 been such that the error of the forecasts since 1911 is less than the error which we obtain when the official formula is applied to the data

TABLE 21. — Error of the Forecasts of the Yield Per Acre of Cotton, by Means of the Official Formula: (1) For the Years 1894–1910; (2) For the Years 1911–1914; (3) For the Years 1907–1910

Representative states	Error for the years 1894–1910, S'_1 Error for the years 1911–1914, S'_2 Error for the years 1907–1910, S'_3														
	May			June			July			August			September		
	S'_1	S'_2	S'_3	S'_1	S'_2	S'_3	S'_1	S'_2	S'_3	S'_1	S'_2	S'_3	S'_1	S'_2	S'_3
Texas	27	23	22	23	19	22	19	19	17	19	13	15	15	8	12
Georgia	17	20	10	15	15	6	12	12	4	11	15	4	11	13	3
Alabama	18	14	18	15	7	11	13	10	8	11	13	5	10	10	5
South Carolina	22	23	9	20	16	8	18	13	9	19	19	8	15	16	9

from 1894 to 1911? (b) Is it true that the year 1911 begins a period of marked improvement in the crop-reporting service; or, in more definite terms, is the error of the forecasts for the years since 1911 less than the error for the same length of time preceding 1911?

Table 21 supplies the material for a decision. We recall that S' is the coefficient that measures the error of the forecasts for the whole period 1894–1914. S'_1, S'_2, S'_3, in Table 21, have the following meanings: S'_1 is the error of the forecasts when the official formula is applied to the data for the 17 years 1894–1910; S'_2 is the error of the forecasts for the years 1911–1914; S'_3 is the error for the preceding four years, 1907–1910. We shall consider first the comparative values of S'_1, S'_2.

In Texas, the forecasts from the data for May, June, August, and September show that S'_1 is in all four months greater than S'_2; for July, the two coefficients are equal. Taking the values of these coefficients as they are, without regard to their probable errors, it is legitimate to infer that in Texas, during the years 1911–1914 as compared with the 17 years 1894–1910, there has been an improvement in the crop-reporting service.

With regard to Georgia the contrary inference must be made. In three out of five cases S'_1 is less than S'_2, while in the remaining two cases the coefficients are equal.

In Alabama, there has possibly been some improvement. In three out of five cases S'_1 is greater than S'_2; in one case S'_1 is less than S'_2; and in one case the two coefficients are equal.

In South Carolina there has been no change. In

two out of five cases S_1' is less than S_2'; in two cases S_1' is greater than S_2'; and in one case they are equal.

From these findings the conclusion may be drawn that, as far as the representative states are concerned, there has been no such improvement in the crop-reporting service during the 21 years, 1894–1914, as to make questionable the testing of the accuracy of the official forecasting formula by its application to the data which we have actually employed.

We come now to the other question of fact: Is it true that the year 1911, when the Department of Agriculture published its forecasting formula, initiated a period of a more accurate crop-reporting service? The comparative values of S_2', S_3' give the necessary figures. S_2' measures the accuracy of the prediction for the four years 1911–1914, and S_3' measures the accuracy of the forecasts when the official formula is applied to the data of the four years preceding 1911, namely, to the years 1907–1910. There are 20 cases in which the values of S_2', S_3' may be compared, and we find, to our surprise, that in 15 out of 20 possible cases S_3' is less than S_2'; that is to say, there has been absolutely no improvement in the recent crop-reporting service.

It cannot reasonably be maintained, therefore, that because of improvements in the official forecasting service, the inferences which we have drawn, concerning the comparative accuracy of the forecasts from the weather and the official forecasts from the laboriously collected data about crop conditions, have not an abiding value.

TABLE 22. — Texas. Degrees of Correlation Between the Yield Per Acre of Cotton and Temperature and Rainfall, when Different Methods of Computation Are Used

The variables	Values of the Coefficient of Correlation									
	May		June		July		August		September	
	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall
Yield-ratio, temperature-ratio, rainfall-ratio	.163	.056	—.047	.195	—.474	—.057	—.628	.691	.133	—.078
First differences of yield per acre, temperature, rainfall	.151	.157	—.080	.161	—.529	.109	—.712	.695	—.123	.110
Second differences of yield per acre, temperature, rainfall	.206	.123	—.055	.103	—.577	.204	—.770	.789	—.231	.244

(1) The yield-ratios, temperature-ratios, and rainfall-ratios used in computing the coefficients of correlation were for the 21 years 1894–1914, except that in case of the rainfall-ratios for August, the record for 1914 was omitted.

(2) The first and second differences were based upon the raw data from 1890 to 1914.

(3) The correlation between the second differences of yield per acre and (a) the second differences of July temperature, (b) the second differences of August temperature, and (c) the second differences of August rainfall is $R = .887$.

Table 22 presents important calculations bearing upon the adequacy of the method which we have employed to measure the relation between the yield per acre of cotton and the variations in temperature and in rainfall. As has been already suggested, the statistics of the yield per acre of cotton and the records of temperature and rainfall are the summary expressions covering results of three classes of changes — secular, cyclical, and random changes — with three different sets of causes. Moreover the series which we have made the basis of our investigation cover only a quarter of a century. Our problem is to discover the true relations between the weather and the yield, to the end that the knowledge of the natural relations may be made a basis of a reliable system of forecasting the yield. But we are at once confronted with a dilemma. If we employ the best method for measuring the true relation between two variables each of which is expressed in a numerical series made up of secular, cyclical, and random changes, we need very long series of observations, whereas our own particular series are relatively short; if, on the other hand, we apply a simpler method to our limited but complex series, our findings are liable to be spurious in the sense of resulting from the inadequacy of the method which we have employed and not resting upon natural relations of the variables. The best we can do under the circumstances is to compare the results of applying different methods to the same problem, and if we find that with the most approved devices there is an agreement in the essential results, we are justified in an accession of faith in our work.

The query that naturally presents itself with refer-

ence to the method which we have adopted, in this chapter, of correlating ratios – the ratios being derived from progressive averages of three years — is whether the correlations that we obtain are true correlations in the sense of indicating the presence of real causes, and not spurious correlations having their origin in some singularity of the method itself. Until very recently statistical science offered no means of treating adequately the difficulty that confronts us, and, indeed, where the series to be compared are relatively short, there is even now no entirely satisfactory method for ascertaining the true relation of the variables. But the *Variate Difference Correlation Method* which has been gradually elaborated by Professor Pearson[1] and his co-workers, is, as far as I am aware, the method that is freest from theoretical objections when it is applied to such limited series as we are compelled to work with. If, therefore, a comparison of the results of the method of correlating ratios with the results of the application of the *Variate Difference Method* shows a substantial agreement in the signs and magnitudes of the coefficients, then the force of one of the chief objections to our work is greatly reduced.

In order to make a test case, we shall take the correlation of the yield per acre of cotton and the records of temperature and rainfall in Texas. The reason for selecting Texas is because we have found that in Texas alone, among the representative states, the forecasts from the weather are not in all cases better than the forecasts from the condition of the crop. In Texas, only

[1] *Biometrika*, Vol. X, Parts II and III, November, 1914, pp. 340–355, and the references there cited.

two of the five months give forecasts from the weather that are better than the forecasts from the condition of the crop, while in all of the other states, for every month, the forecast from the weather is better than the forecast from the condition of the crop. A comparison, therefore, of the results in Texas, in the manner which we have proposed, will have the advantage not only of a test of our method but will also settle the question whether, by the use of a different method, we might not obtain for Texas the uniformly better forecasts from the weather which we have obtained for the other representative states.

An examination of the computations in Table 22 shows that, when attention is given to the probable errors of the coefficients, there is a remarkable agreement between the correlations of ratios and the correlations of first differences and of second differences. Two essential points are brought out by the three rows of coefficients: (1) All three series indicate that the critical factors in the growth season of cotton, in Texas, are the July temperature, the August temperature, and the August rainfall; (2) The method that we have adopted in this chapter to measure the relation between cotton yield and the elements of the weather does not, in this test case, exaggerate the degree of association between the variables.

From the results of this test case we draw the important conclusion that the method of correlating ratios does enable us to discover the critical factors in the growth of the cotton plant; and that the forecasts based upon the correlations of ratios are not spurious, but rest upon real causes.

APPENDIX

TABLE 23. — Texas. Official Monthly Reports on the Condition of the Growing Cotton Crop, and Official Final Estimates of the Annual Yield Per Acre in Pounds of Cotton Lint

Year	Condition of the Crop					Yield per Acre in Pounds of Lint
	June 1	July 1	August 1	September 1	October 1	
1889	95	90	91	81	78	169
1890	84	89	82	77	77	196
1	91	95	92	82	78	195
2	81	87	86	81	77	291
3	82	84	72	63	65	151
4	94	99	85	84	88	235
5	79	76	71	56	58	151
6	92	80	69	62	57	104
7	87	88	78	70	64	165
8	89	92	91	75	73	212
9	90	93	87	61	56	185
1900	71	78	83	77	78	226
1	84	86	74	56	51	159
2	95	73	77	53	47	148
3	70	79	82	76	54	143
4	84	89	91	77	69	183
5	69	72	71	70	69	164
6	87	82	86	78	74	225
7	70	72	75	67	60	130
8	77	80	82	75	71	196
9	78	79	70	59	52	125
1910	83	84	82	69	63	145
11	88	85	86	68	71	186
12	86	89	84	76	75	206
13	84	86	81	64	63	150
14	65	74	71	79	70	184

TABLE 24. — GEORGIA. OFFICIAL MONTHLY REPORTS ON THE CONDITION OF THE GROWING COTTON CROP, AND OFFICIAL FINAL ESTIMATES OF THE ANNUAL YIELD PER ACRE IN POUNDS OF COTTON LINT

Year	Condition of the Crop					Yield per Acre in Pounds of Lint
	June 1	July 1	August 1	September 1	October 1	
1889	80	86	91	90	87	155
1890	94	95	94	86	82	165
1	80	85	86	82	78	155
2	87	88	84	79	75	160
3	87	86	83	77	76	136
4	76	78	85	84	79	155
5	82	88	87	76	72	152
6	95	94	92	71	67	122
7	84	85	95	80	70	178
8	89	90	91	80	75	183
9	88	85	79	69	64	159
1900	89	74	77	69	67	172
1	80	72	78	81	73	167
2	94	91	83	68	62	165
3	75	75	77	81	68	158
4	78	85	91	86	78	205
5	84	82	82	77	76	200
6	86	82	74	72	68	165
7	74	78	81	81	76	190
8	80	83	85	77	71	190
9	84	79	78	73	71	184
1910	81	78	70	71	68	173
11	92	94	95	81	79	240
12	74	72	68	70	65	163
13	69	74	76	76	72	208
14	80	83	82	81	81	239

TABLE 25. — Alabama. Official Monthly Reports on the Condition of the Growing Cotton Crop, and Official Final Estimates of the Annual Yield Per Acre in Pounds of Cotton Lint

Year	Condition of the Crop					Yield per Acre in Pounds of Lint
	June 1	July 1	August 1	September 1	October 1	
1889	83	87	90	91	87	163
1890	93	95	93	84	80	160
1	89	87	89	83	76	165
2	91	90	83	72	69	135
3	82	80	79	78	76	148
4	88	87	94	86	84	160
5	85	83	81	71	70	135
6	103	98	93	66	61	124
7	81	85	88	80	73	155
8	89	91	95	80	76	195
9	86	88	82	76	70	176
1900	87	70	67	64	62	151
1	76	80	82	75	65	156
2	92	84	77	54	52	144
3	73	76	79	84	68	161
4	80	85	90	84	76	182
5	87	83	79	70	70	173
6	81	84	83	76	68	165
7	65	68	72	73	68	169
8	78	82	85	77	70	179
9	83	64	68	66	62	142
1910	83	81	71	72	67	160
11	91	93	94	80	73	204
12	74	76	73	75	68	173
13	75	79	79	72	67	190
14	85	88	81	77	78	209

TABLE 26. — South Carolina. Official Monthly Reports on the Condition of the Growing Cotton Crop, and Official Final Estimates of the Annual Yield Per Acre in Pounds of Cotton Lint

Year	Condition of the Crop					Yield per Acre in Pounds of Lint
	June 1	July 1	August 1	September 1	October 1	
1889	78	84	90	87	81	141
1890	97	95	95	87	83	175
1	80	80	83	81	72	160
2	91	94	83	77	73	184
3	88	83	75	63	62	142
4	83	88	95	86	79	168
5	72	84	81	82	64	141
6	97	98	88	70	67	129
7	87	86	92	84	74	189
8	85	90	89	81	79	245
9	86	88	78	66	62	165
1900	85	79	74	60	57	167
1	80	70	75	80	67	141
2	97	95	88	74	68	199
3	76	74	76	80	70	178
4	81	88	91	87	81	215
5	78	78	79	75	74	220
6	82	77	72	71	66	175
7	77	79	81	83	77	215
8	81	84	84	76	68	219
9	83	77	77	74	70	210
1910	78	75	70	73	70	216
11	80	84	86	74	73	280
12	83	79	75	73	68	209
13	68	73	75	77	71	235
14	72	81	79	77	72	255

TABLE 27. — TEXAS. TEMPERATURE (DEGREES FAHRENHEIT) AND RAINFALL (INCHES) IN EASTERN AND CENTRAL TEXAS

Year	May		June		July		August		September	
	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall
1891	71.3	2.64	81.7	2.48	82.7	1.71	81.1	1.70	77.3	2.26
2	73.5	4.55	79.5	4.37	82.8	1.85	80.9	4.12	75.5	1.40
3	73.0	4.71	80.0	2.88	85.0	0.75	81.6	2.19	79.3	1.79
4	74.5	3.60	78.8	2.56	82.0	2.03	79.8	5.26	76.9	2.57
5	71.3	7.01	78.6	6.18	82.8	4.11	83.5	1.78	80.5	1.79
6	77.8	1.62	83.4	0.99	84.8	1.94	85.2	1.64	78.0	4.56
7	72.3	4.33	80.6	3.65	85.9	1.26	83.0	2.32	76.9	2.58
8	74.5	3.55	80.2	5.63	82.3	2.09	82.6	2.50	77.7	1.61
9	77.1	3.49	79.8	6.56	82.5	2.00	86.6	0.36	76.7	1.16
1900	72.3	5.89	81.8	1.85	81.8	4.50	82.0	2.98	80.8	6.60
1	72.4	4.22	81.8	1.08	85.4	1.99	85.2	1.77	76.8	3.27
2	76.4	3.78	82.8	2.19	81.6	7.73	85.2	0.11	75.0	4.51
3	70.0	2.27	73.9	3.70	81.3	5.91	82.4	1.71	74.7	2.98
4	72.3	4.76	79.4	4.57	81.8	2.45	81.9	2.16	79.0	2.82
5	74.9	5.95	81.1	4.47	80.7	4.90	83.9	1.03	79.5	2.02
6	72.6	3.96	80.5	3.83	80.8	5.13	80.8	4.46	78.3	3.50
7	67.6	6.80	80.4	1.85	83.1	2.87	85.1	1.01	79.0	1.29
8	73.2	7.87	81.3	2.19	81.8	2.60	82.2	2.28	76.0	3.54
9	72.2	2.91	81.1	3.07	86.5	1.56	85.4	2.00	78.0	0.88
1910	71.4	4.04	80.2	1.79	84.8	1.15	86.4	0.83	81.5	1.78
11	73.2	1.50	84.7	0.65	82.8	4.36	84.6	2.93	83.1	1.53
12	74.0	2.21	77.4	3.28	85.4	1.04	84.0	3.47	77.9	0.77
13	73.0	3.55	79.0	2.66	84.9	1.52	84.7	1.03	73.2	5.70
14	71.1	7.81	82.4	1.31	86.2	0.91	80.7	8.95	77.4	1.39

TABLE 28. — GEORGIA. TEMPERATURE (DEGREES FAHRENHEIT) AND RAINFALL (INCHES)

Year	May		June		July		August		September	
	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall
1892	72.1	2.16	78.1	5.93	78.1	6.13	76.3	6.18	72.4	3.61
3	70.5	2.04	74.9	4.53	81.2	2.78	78.3	6.63	75.0	2.86
4	71.4	2 51	77.1	2.72	78.1	7.82	78.5	5.20	75.1	3.72
5	69 8	4.20	77.9	3.93	79.0	4.96	79.1	7.55	77.0	1.53
6	76.0	2.54	77.7	3.55	80.0	8.26	81.6	2.89	76.3	2.37
7	69.3	1.52	80.8	3.51	80.6	5.74	78.3	5.07	73.8	2.83
8	74.0	1 12	80.1	3 27	79.8	8.14	78.5	10.09	75.3	4.76
9	75.2	1 76	80.2	2.59	80.2	4.53	81.1	4.58	73.4	1.40
1900	70.8	2.46	75 6	8.98	80.3	5.12	82 3	2.55	77.3	3.06
1	71 4	5 71	77.6	5.26	81.5	4.18	78.2	9.92	73.1	5.19
2	75.5	2 34	79 5	3.54	82.0	4.55	80.3	3.92	73.0	4.67
3	70.3	5 47	74 4	6.00	80.0	4.00	80.7	5 56	73 1	4 40
4	70.2	2.23	77.4	2.95	79.0	3.81	77.6	7.33	76 0	1 48
5	74.5	5.02	78.7	3.69	80.3	5.57	78.6	4.96	76.8	2.95
6	70.2	4.32	77.8	6.31	77.8	8.41	80.1	5.82	77.5	5 29
7	70.1	4 26	76.0	4.29	81 0	5.04	79.5	4.10	75.5	6.24
8	72.1	2.67	77.5	3.40	79.8	5.27	79.0	6.04	73.2	3.11
9	69.7	4.43	78.5	5.48	79.0	5.25	79.8	4.53	74 2	3.17
1910	69.8	3.61	75.3	7.16	78.9	5.68	79.1	3.68	76 3	2 48
11	73.1	2 14	80.9	2.78	78.3	5.44	79.4	6.18	79 6	2.95
12	72.6	4.08	75.4	6 83	79.5	5.71	79.1	4.91	77.3	5.73
13	71.9	2.27	76.4	4.83	81 3	5.61	79.3	4.03	72.3	4.12
14	72.7	0.74	82.2	3.51	80.8	4.74	79.1	5.99	72.7	3.53

TABLE 29. — ALABAMA. TEMPERATURE (DEGREES FAHRENHEIT) AND RAINFALL (INCHES)

Year	May		June		July		August		September	
	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall
1896	75.8	3.44	77.2	5.24	80.9	5.09	82.2	2.30	75.8	1.76
7	68.6	1.56	80.9	1.85	81.1	4.78	78.8	5.58	75.6	0.55
8	73.6	0.82	80.4	3.60	80.0	6.06	78.7	7.43	75.5	3.58
9	75.9	2.03	79.8	2.54	80.4	6.76	81.3	3.68	72.7	0.66
1900	71.2	2.64	76.4	11.08	79.8	4.93	81.6	2.89	77.8	4.00
1	69.8	5.08	78.5	2.80	82.2	3.40	78.6	8.86	72.1	4.19
2	75.4	2.34	80.8	1.28	82.8	2.50	82.1	3.48	73.4	4.28
3	69.6	6.05	73.2	4.88	80.0	3.98	80.5	3.57	73.2	1.42
4	69.6	2.98	77.8	2.94	79.6	4.80	78.4	5.55	76.8	1.36
5	74.2	5.51	79.0	4.56	79.4	4.56	79.2	5.30	76.2	2.51
6	69.7	4.63	78.9	3.45	78.8	8.50	80.4	3.78	78.2	8.44
7	68.0	7.94	75.6	2.85	81.0	5.00	80.4	3.50	74.8	5.50
8	71.4	5.34	77.5	2.75	79.8	4.72	79.4	3.44	74.2	2.42
9	68.6	6.51	78.0	7.82	79.3	4.52	81.0	3.30	73.7	2.87
1910	68.9	3.86	75.6	6.98	78.6	7.18	79.7	2.73	77.5	2.21
11	72.9	2.85	80.6	3.86	78.0	5.66	79.1	4.97	80.4	2.32
12	72.0	3.60	75.1	5.10	79.7	5.17	79.2	5.68	77.1	4.79
13	71.6	3.14	77.5	3.54	81.1	5.00	80.5	2.58	73.1	6.96
14	71.8	1.05	83.1	2.66	81.6	4.23	79.1	6.41	72.4	4.69
15	74.5	6.34	78.8	3.66	80.4	5.23	78.9	5.07	76.6	4.43

TABLE 30. — South Carolina. Temperature (Degrees Fahrenheit) and Rainfall (Inches)

Year	May		June		July		August		September	
	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rainfall	Temperature	Rain fall	Temperature	Rainfall
1891	69.0	3.57	79.3	3.20	77.9	5.95	78.9	8.79	74.3	2.66
2	71.1	5.53	75.1	5.25	78.9	7.44	79.5	4.42	72.5	6.41
3	70.2	4.13	76.5	7.64	82.0	3.87	77.5	12.45	74.7	4.42
4	70.7	3.43	77.0	3.91	77.7	8.24	77.9	7.28	75.0	6.51
5	69.0	4.36	78.2	3.04	79.5	4.17	79.4	7.95	76.9	1.29
6	76.7	2.74	77.9	5.42	80.7	8.17	80.4	4.14	75.0	2.94
7	69.3	2.39	79.2	5.44	80.2	5.01	78.0	5.16	73.3	2.91
8	73.8	1.35	79.7	4.15	80.0	7.81	78.7	9.81	76.0	4.06
9	73.7	1.68	79.4	3.89	80.0	4.03	81.2	6.26	72.8	2.55
1900	70.2	2.37	76.2	7.94	81.2	4.08	83.0	2.13	77.1	2.83
1	71.4	7.31	76.7	6.55	81.4	4.52	78.6	9.01	73.1	4.66
2	74.0	2.69	78.5	4.48	80.8	3.79	78.6	5.07	72.1	3.74
3	70.7	2.69	74.2	8.09	80.4	3.59	80.6	7.15	72.7	3.62
4	70.6	2.04	77.0	4.06	79.4	5.96	77.6	8.47	75.8	2.46
5	73.4	5.70	78.9	1.92	80.4	6.16	77.9	5.69	76.2	1.91
6	70.7	3.00	78.4	8.88	78.4	8.40	80.6	6.62	78.0	4.85
7	70.8	4.51	75.8	5.92	81.4	5.06	79.4	5.41	76.0	5.91
8	71.8	2.92	76.6	4.90	79.8	5.43	78.6	9.11	72.4	2.86
9	69.5	4.26	79 2	6 87	78.6	4.92	78.8	4.83	72 0	3.74
1910	69.8	4.03	75.5	7.78	79.4	5.83	79.0	6.00	75.8	3.10
11	72.6	0.65	80.9	3.42	79.8	3.79	80.0	6.05	78.9	3.33
12	72.4	4.08	75.5	5.68	79.6	5.22	79.2	3.69	77.7	5.91
13	71.7	2.13	76.2	5.53	81.9	4.78	78.9	3.76	71.7	4.66
14	72.1	0.83	81.1	3.80	80.0	5.56	79.0	5.88	71.3	3.63

CHAPTER V

THE LAW OF DEMAND FOR COTTON

"There is a general agreement as to the character and directions of the changes which various economic forces *tend* to produce. . . . Much less progress has been made towards the *quantitative* determination of the *relative strength* of different economic forces."

— ALFRED MARSHALL.

THE investigations of the preceding two chapters have made us acquainted with the degree of reliability of the Government reports on the prospective cotton crop and with the measure of accuracy with which, at any stage in the growth season, the prospective yield of cotton may be calculated from the past conditions and vicissitudes of the weather. The new problem that we face in this chapter carries the inquiry to its final stage: Assuming that the ultimate volume of the crop may be forecast with a known degree of precision, is it possible to predict the relation that will subsist between the size of the crop and the price of cotton lint? Is it possible to know the dynamic law of the demand for cotton?

Two Practical Methods of Approach

In Chapters III and IV, we found that the method of progressive averages enabled us to get valuable results in the problem of forecasting the amount of production from the Government reports on the condition of the growing crop, and from the records of temperature and rainfall in the Cotton Belt. We shall test the helpful-

ness of this same device in our present inquiry as to the form of the concrete law of demand for cotton.

Method of progressive averages. In Table 31 the data[1] are collected for computing the relation between the price-ratio and the production-ratio of cotton. The problem to be solved may be put into symbolic form: Let p be the mean price per pound of cotton for any given year, and $\bar{p}_3$ be the mean price for the preceding three years; let P be the total production of cotton for the given year, and $\bar{P}_3$ be the mean production for the preceding three years. Our problem is to find (1) the coefficient of correlation measuring the relation between $p/\bar{p}_3$ and $P/\bar{P}_3$; (2) the statistical law connecting $p/\bar{p}_3$ with $P/\bar{P}_3$, which is the concrete law of demand for cotton; (3) the error incurred in using the law of demand for cotton as a formula with which to forecast the price of cotton from the prospective size of the crop.

The values of the series $p/\bar{p}_3$ and $P/\bar{P}_3$, for the period 1890 to 1913, are given in columns 4 and 7 of Table 31. The calculation of the items that constitute the solution of our problem gives:

(1) The coefficient of correlation between $p/\bar{p}_3$ and $P/\bar{P}_3$ is $r = -.706$;

[1] The crude data are taken from the *Statistical Abstract of the United States*, 1914, p. 505. "The production statistics relate, when possible, to the year of growth, but when figures for the year are wanting, a commercial crop which represents the trade movement is taken. The statistics of production have been compiled from publications of the United States Department of Agriculture for 1860 to 1898. Census figures have, however, been used when available, including those for 1899 to date." *Ibid.*, note 1.

"The value of lint per pound shown since 1902 relates to the average grade of upland cotton marketed prior to April 1 of the following year; from 1890 to 1901, the average price of middling cotton on the New Orleans Cotton Exchange." *Ibid.*, note 2.

TABLE 31. — THE PRODUCTION-RATIO AND THE PRICE-RATIO OF COTTON

Year	Equivalent 500 Pound Bales, Gross Weight (Millions of Bales) P	Mean Production for the Preceding Three Years $\bar{P}_3$	Production-Ratio $P/\bar{P}_3$	Price per Pound Upland Cotton (Cents) p	Mean Price for the Preceding Three Years $\bar{p}_3$	Price-Ratio $p/\bar{p}_3$
1887	6.88			10.3		
8	6.92			10.7		
9	7.47			11.5		
1890	8.56	7.09	120.7	8.6	10.8	79.6
1	8.94	7.65	116.9	7.3	10.3	70.9
2	6.66	8.32	80.0	8.4	9.1	92.3
3	7.43	8.05	92.3	7.5	8.1	92.6
4	10.03	7.68	130.6	5.9	7.7	76.6
5	7.15	8.04	88.9	8.2	7.3	112.3
6	8.52	8.20	103.9	7.3	7.2	101.4
7	10.99	8.57	128.2	5.6	7.1	78.9
8	11.44	8.89	128.7	4.9	7.0	70.0
9	9.35	10.32	90.6	7.6	5.9	128.8
1900	10.12	10.59	95.6	9.3	6.0	155.0
1	9.51	10.30	92.3	8.1	7.3	111.0
2	10.63	9.66	110.0	8.2	8.3	98.8
3	9.85	10.09	97.6	12.2	8.5	143.5
4	13.44	10.00	134.4	8.7	9.5	91.6
5	10.58	11.31	93.5	10.9	9.7	112.4
6	13.27	11.29	117.5	10.0	10.6	94.3
7	11.11	12.43	89.4	11.5	9.9	116.2
8	13.24	11.65	113.6	9.2	10.8	85.2
9	10.00	12.54	79.7	14.3	10.2	140.2
1910	11.61	11.45	101.4	14.7	11.7	125.6
11	15.69	11.62	135.0	9.7	12.7	76.4
12	13.70	12.43	110.2	12.0	12.9	93.0
13	14.16	13.67	103.6	13.1	12.1	108.3

(2) The concrete law of demand for cotton is $y = -.975x + 206.03$; where x is put for $P/\bar{P}_3$, and y is the most probable value of $p/\bar{p}_3$, corresponding to the given value of $P/\bar{P}_3$;

(3) The accuracy with which the law of demand for cotton may be used to forecast the price of cotton lint is measured by $S = \sigma_y\sqrt{1 - r^2} = 16.38$.

Method of percentage changes.[1] In Table 32 the crude statistical data of production and prices are utilized in a different way. From year to year both the price and the production of cotton undergo changes, and in the construction of Table 32 the hypothesis in mind suggested that there is a close relation between the percentage change of the price in any given year over the price of the preceding year, and the percentage change in production of the given year over the production of the preceding year. The percentage changes are tabulated in columns 4 and 7.

The calculations based upon the data of this Table show that

(1) The coefficient of correlation between the percentage change in price and the percentage change in production is $r = -.819$;

(2) The dynamic law of demand for cotton is $y = -1.08x + 8.81$; where x is put for the percentage change in production, and y is the most probable value of the percentage change in price, corresponding to the given percentage change in production;

(3) The accuracy with which the dynamic law of

[1] A more ample description of this method is contained in *Economic Cycles: Their Law and Cause*, Chapter IV.

demand for cotton may be used to forecast the percentage change in the price of cotton lint is measured by $S = \sigma_y\sqrt{1 - r^2} = 15.18$.

TABLE 32. — PERCENTAGE CHANGES IN THE PRICE AND PRODUCTION OF COTTON LINT

Year	Equivalent 500 Pound Bales, Gross Weight (Millions of Bales)	Change over the Preceding Year	Percentage Change over the Preceding Year	Price per Pound Upland Cotton (Cents)	Change over the Preceding Year	Percentage Change over the Preceding Year
1889	7.47			11.5		
1890	8.56	+1.09	+14.59	8.6	−2.9	−25.22
1	8.94	+0.38	+4.44	7.3	−1.3	−15.12
2	6.66	−2.28	−25.50	8.4	+1.1	+15.07
3	7.43	+0.77	+11.56	7.5	−0.9	−10.71
4	10.03	+2.60	+34.99	5.9	−1.6	−21.33
5	7.15	−2.88	−28.71	8.2	+2.3	+38.98
6	8.52	+1.37	+19.16	7.3	−0.9	−10.98
7	10.99	+2.47	+28.99	5.6	−1.7	−23.29
8	11.44	+0.45	+4.10	4.9	−0.7	−12.50
9	9.35	−2.09	−18.27	7.6	+2.7	+55.10
1900	10.12	+0.77	+8.24	9.3	+1.7	+22.37
1	9.51	−0.61	−6.03	8.1	−1.2	−12.90
2	10.63	+1.12	+11.78	8.2	+0.1	+1.23
3	9.85	−0.78	−7.34	12.2	+4.0	+48.78
4	13.44	+3.59	+36.45	8.7	−3.5	−28.69
5	10.58	−2.86	−21.28	10.9	+2.2	+25.29
6	13.27	+2.69	+25.43	10.0	−0.9	−8.26
7	11.11	−2.16	−16.28	11.5	+1.5	+15.00
8	13.24	+2.13	+19.17	9.2	−2.3	−20.00
9	10.00	−3.24	−24.47	14.3	+5.1	+55.43
1910	11.61	+1.61	+16.10	14.7	+0.4	+2.80
11	15.69	+4.08	+35.14	9.7	−5.0	−34.01
12	13.70	−1.99	−12.68	12.0	+2.3	+23.71
13	14.16	+0.46	+3.36	13.1	+1.1	+9.17

Figure 11 makes clear to the eye the measure of agreement between the actual percentage changes in price and the percentage changes as they are predicted from the law of demand.

A comparison of the results we obtain from these two methods of deriving the law of demand for cotton shows that there is very little difference between them so far as the accuracy of the forecasts are concerned.

But, in Chapter I, we have said that it is possible to forecast the price of cotton from the size of the crop with greater accuracy than the *Bureau of Statistics* can forecast the yield of cotton from the known condition of the growing crop. This statement we shall now prove. Throughout our investigations we have measured the accuracy of forecasts by $S = \sigma_y\sqrt{1-r^2}$, where σ_y is the standard deviation of a concrete series, and r is the correlation between two series. S measures the accuracy of the forecasts because it shows how the prediction formula enables one to reduce their variability. If there were no forecasting formula the variability of the series that we wish to know would be σ_y, but by the use of the formula the variability of the forecasts is only $\sigma_y\sqrt{1-r^2}$. The factor $\sqrt{1-r^2}$ measures the reduction in variability that is gained by means of the forecasting formula. If, therefore, we wish to compare the accuracy of forecasts of two different series, the measure of the relative accuracy is given by $\sqrt{1-r^2}$, and the smaller the value of $\sqrt{1-r^2}$, the greater the accuracy of the forecasts. The same idea may be put in a different way by saying that the greater the value of r, the greater the accuracy of the forecasts.

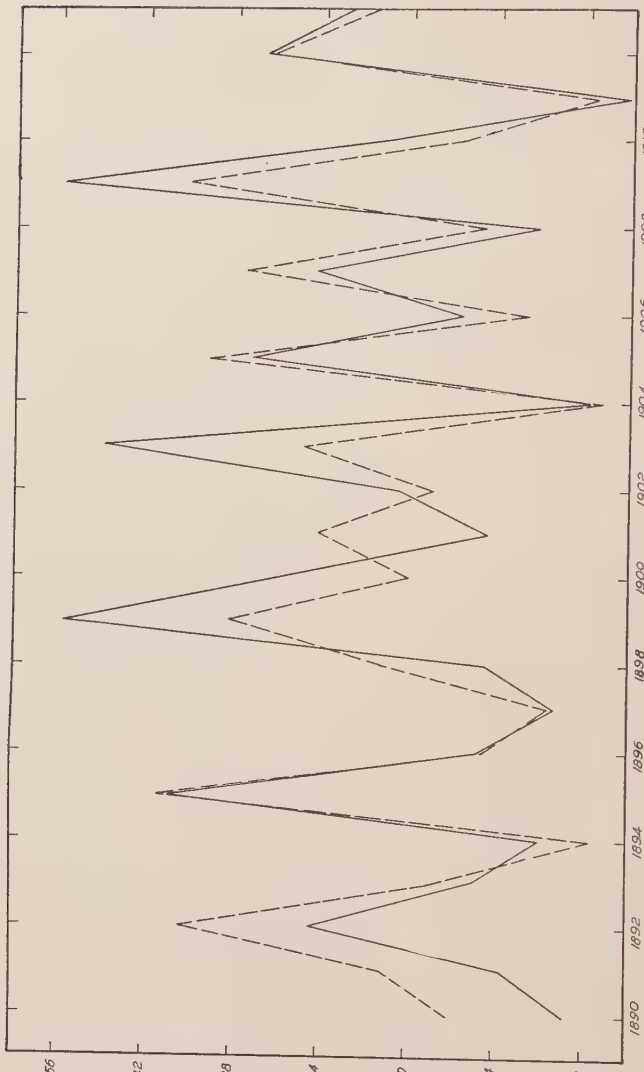

FIGURE 11. — The actual percentage changes in the prices of cotton, —, and the percentage changes as they are predicted, – –, by means of the dynamic law of demand for cotton, $y = -1.08x + 8.81$.

If now we refer to Chapter III, "The Government Crop Reports," we find that the correlation between the predicted yield-ratio of cotton and the actual yield-ratio is, for the month of May, $r = -.049$; for June, $r = .292$; for July, $r = .595$; for August, $r = .576$; for September, $r = .685$. But the calculations of the present chapter have shown that the correlation between the price-ratio of cotton lint and the production-ratio is $r = -.706$; and the correlation between the percentage change of prices and the percentage change of production is $r = -.819$.

Statics and Dynamics Discriminated

The law of demand for cotton, in the form in which it was treated in the preceding section, was deduced from official descriptions of changes extending over a quarter of a century. We have found that the correlation between the percentage change of prices and the percentage change of production is $r = -.819$; if we had used the data extending as far back as 1869, when the disastrous effects of the Civil War upon prices had not yet been overcome, we should have found $r = -.736$, which is still a high coefficient. Our law of demand is a dynamic law; it is a summary description of a routine in concrete affairs.

The law of demand in statical economics is of a different quality, and the hypothetical limitations of deductions based upon it are not always kept in mind.

According to Professor Marshall, "there is . . . one general *law of demand:* The greater the amount to be sold, the smaller must be the price at which it is offered

in order that it may find purchasers; or, in other words, the amount demanded increases with a fall in price, and diminishes with a rise in price." "The one universal rule to which the demand curve conforms is that it is *inclined negatively* throughout the whole of its length." [1] This statement of the law is absolute. Remembering the explicit claim made in the *Preface* to the last edition of Professor Marshall's work (the fifth edition of 1907) that his volume is "concerned throughout with forces that cause movement: and its key-note is that of dynamics," one might fall into the error of supposing that the deductions based upon the law applied directly to actual phenomena. But Professor Marshall has carefully pointed out the reservations that are made:

(1) It is assumed in giving definite form to the law of demand for any one commodity that the prices of all other commodities remain constant. This is the usual *cæteris paribus* assumption.[2] With regard to this hypothesis I should like to quote the comment of Professor Marshall's sympathetic fellow-worker, Professor Edgeworth: "Demand curves as usually understood involve a postulate which is frequently not fulfilled; namely, that while the price of the article under consideration is varied, the prices of all other articles remain constant. This postulate fails in the case of rival commodities such as beef and mutton. The price of one of these cannot be supposed to rise or fall considerably without the price of the other being affected. The same is true of commodities for which there is a joint-demand as for malt and hops. And in case of a

[1] Marshall: *Principles of Economics*, 5th ed., p. 99, note 2.
[2] *Ibidem*, pp. x, 100.

necessary of life the price cannot be supposed to increase indefinitely without the prices of other articles falling, owing to the retrenchment of expenditure on articles other than necessaries." "It is true, indeed, that the postulate which has been stated might be dispensed with. But this can only be done at the sacrifice of two of the characteristic advantages which demand curves offer the theorist. First, unless this postulate is granted, it is hardly conceivable that, when the prices of several articles are disturbed concurrently, the collective demand curve may be predicted by ascertaining the disposition of the individual — a conception which aids us to apprehend the working of a market. Secondly, when the prices of all commodities but one are not supposed fixed, there no longer exists that exact correlation between the demand curve and the interests of consumers in low prices which Prof. Marshall has formulated as 'consumers' rent.'" [1]

(2) The validity of the law is limited to a point in time.[2] Refering to this limitation, Professor Edgeworth remarks: "There is an artificial rigidity in demand curves which imperfectly correspond to the flux character of human desires. One cause of change is the formation of new habits. The increased use of petroleum is not to be ascribed simply to the fall in price, the demand curve being supposed constant, but rather to the fact that 'petroleum and petroleum lamps have become familiar to all classes of society' (Marshall)." [3]

[1] Palgrave's *Dictionary of Political Economy*, Vol. I, "Demand Curves," pp. 543–544.

Principles of Economics, pp. 94, 100.

[3] Palgrave's *Dictionary of Political Economy*, Vol. I, "Demand Curves," p. 544.

(3) The usual statement of the law of demand does not take "account of the fact that, the more a person spends on anything the less power he retains of purchasing more of it or of other things, and the greater is the value of money to him (in technical language every fresh expenditure increases the marginal value of money to him)." [1]

(4) When, however, account is taken of the varying marginal utility of money, it is possible that the demand curve for food, on the part of the "poorer labouring families," shall be positively inclined. But in the statement of the law of demand, Professor Marshall has said "the one universal rule to which the demand curve conforms is that it is inclined negatively." [2]

(5) "Again, the demand for a commodity on the part of dealers who buy it only with the purpose of selling it again, though governed by the demand of the

[1] *Principles of Economics*, p. 132.

[2] *Ibidem*, pp. 132, 99 note.

Referring to Professor Pareto's recognition of the theoretical possibility of obtaining curves of demand of the positive type, M. Zawadzki makes the extremely valuable criticism which I have italicized in the following quotation: "Quelle est la valeur d'une telle conclusion? N'est-elle pas en contradiction flagrante avec les faits? Il est facile d'imaginer des cas théoriques où la demande diminuerait à la suite d'une diminution de prix. La théorie doit donc être capable d'en tenir compte. *Ont-ils lieu en réalité (et dans l'hypothèse statique) autrement que par exception? Quelle peut être leur importance? Voici des questions, et on pourrait en poser bien d'autres, auxquelles la théorie ne nous répond pas. Dans cet exemple nous touchons, pour ainsi dire, du doigt la puissance et la faiblesse de l'économie mathématique. Nous avons la formule la plus générale, englobant jusqu'à des cas extrêmement rares, mais nous ne pouvons pas en passer aux cas particuliers, pas même distinguer ce qui est l'exception de ce qui est la règle.*" *Les Mathématiques Appliquées à L'Économie Politique*, p. 186.

ultimate consumers in the background, has some peculiarities of its own." [1]

(6) The hope of obtaining concrete, statistical laws of demand was expressed by Jevons in 1871, and has been repeated by Professor Marshall in the successive editions of his Principles from 1890 to 1907. But according to Professor Edgeworth ". . . it may be doubted whether Jevons's hope of constructing demand curves by statistics is capable of realisation." [2]

A Complete Solution of the Problem

The listing of the reservations that are made by Professor Marshall when he states "the one universal rule to which the demand curve conforms" has the double advantage of cautioning his reader against drawing precipitate conclusions as to the applicability to concrete affairs of any theoretical deductions based upon the curve, and of suggesting the imperative need of a more concrete treatment of the law of demand. With the derivation of demand curves from statistical data we shall be concerned in the present section.

We shall be aided in approaching our problem if we put it into symbolic form: Suppose we let x_0 be the percentage change in the price of a commodity, say, for instance, cotton, and let x_1 be the percentage change in the amount of the commodity that is demanded. Then, if my interpretation of Professor Marshall's view is correct, his understanding of the nature of the law of the demand may be described in two stages:

[1] Marshall: *Principles of Economics*, p. 100 n.

[2] Palgrave's *Dictionary of Political Economy*, Vol. I, "Demand Curves," p. 544.

(1) In reality $x_0 = \phi(x_1, x_2, x_3, \ldots x_n)$, where $x_2, x_3, \ldots x_n$ are percentage changes in other factors, some of which are enumerated in Professor Marshall's explicit reservations when he formulates the law of demand. The form of the function ϕ is unknown and the interrelations of $x_1, x_2, x_3, \ldots x_n$ are unknown.

(2) In the statement of the law of demand in its absolute form, x_1 is singled out as the important variable in relation to x_0, and the law of demand in its static form expresses the relation that exists between x_0 and x_1 when $x_2, x_3, \ldots x_n$ are all equal to zero. These variables $x_2, x_3, \ldots x_n$ must be equal to zero since they severally represent *percentage changes*, and the general hypothesis in mind when the static law of demand is formulated is that there shall be no changes in other economic factors.

The misgivings that one feels about conclusions which are based upon the static law of demand are due to the fact that the form of the function ϕ is not known; that the influence of the factors $x_2, x_3, \ldots x_n$ is ignored; and that the interrelations of $x_2, x_3, \ldots x_n$ have not been determined. The misgivings would be removed if these three limitations could be overcome.

The procedure that I wish to introduce — the treatment of the problem statistically by the method of multiple correlation — addresses itself precisely to these three limitations. The ultimate aim of economic theory is to enable us to forecast economic phenomena, and, in this particular problem of the law of demand, we wish to forecast x_0, the percentage change in the price. We know that $x_0 = \phi(x_1, x_2, x_3, \ldots x_n)$, and while we do not know either the form of the function ϕ, or the

interrelations of $x_1, x_2, x_3, \ldots x_n$, the practical problem before us suggests effectual means of overcoming these limitations. For, since our ultimate object is to forecast the value of x_0 and to measure the degree of accuracy with which the forecast is made, we may, with due precautions against spurious results, experiment with different types of the function ϕ and of the interrelations of $x_1, x_2, x_3, \ldots x_n$, and settle upon those types that enable us to forecast x_0 with a degree of precision sufficient for the actual problem in hand.

As a first approximation we naturally take the simplest type of functions. We say, suppose

(1) That the type of the function ϕ is linear, such that

$$\phi(x_1, x_2, x_3, \ldots x_n) = u = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n;$$

(2) That the interrelations of $x_1, x_2, x_3, \ldots x_n$ are also linear, such that, for example $x_1 = b_1 + b_2x_2$. This second supposition will present no difficulties, since we have dealt with the problem of finding the relation between x_1 and x_2 when the connection is of the simple form $x_1 = b_1 + b_2x_2$. With regard to the first supposition, all that we need to do in order to obtain a satisfactory practical solution of our problem is so to determine from the actual statistics the values of the constants $a_0, a_1, \ldots a_n$ that the correlation between x_0 and u, which we designate by R, shall be a maximum. In that case the value of $S = \sigma_0\sqrt{1 - R^2}$, which measures the root-mean-square error of the forecasts by means of the formula $x_0 = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n$, will be a minimum.

Now precisely this problem has already received a

general solution in the statistical theory of multiple correlation. If, for the sake of simplicity, we take the case of three variables, then the equation connecting x_0, x_1, x_2 in such a way that the correlation is a maximum between the actual values of x_0 and the predicted values of x_0 has the form

$$(x_0-\bar{x}_0)=\frac{r_{01}-r_{02}r_{12}}{1-r_{12}^2}\frac{\sigma_0}{\sigma_1}(x_1-\bar{x}_1)+\frac{r_{02}-r_{01}r_{12}}{1-r_{12}^2}\frac{\sigma_0}{\sigma_2}(x_2-\bar{x}_2)$$

$$\text{and } S=\sigma_0\sqrt{1-R^2}, \text{ where } R^2=\frac{r_{01}^2+r_{02}^2-2r_{01}r_{02}r_{12}}{1-r_{12}^2}.$$

The forecasting equation enables us to predict the most probable values of x_0 from the known values of x_1, x_2, and S measures the degree of accuracy with which the forecasts are made.

An example will make this abstract discussion much clearer. Professor Edgeworth makes the statement that,

"One important cause of alteration in demand curves is the increase of the consumer's purchasing power. The case in which that increase is only apparent, being due to a rise in prices (and the converse case), may be specially distinguished. Owing to the variability, it may be doubted whether Jevons's hope of constructing demand curves by statistics is capable of realisation." [1]

Professor Edgeworth doubtless meant that because of the *many* factors tending to produce a variation in the demand schedule it might be doubtful whether Jevons's hope could be realised. But suppose — for

[1] Palgrave's *Dictionary of Political Economy*, "Demand Curves," Vol. I, p. 544.

the sake of simplicity and concreteness but in illustration of a method of complete solution — we limit our inquiry to this question: How may the relation between the price of cotton and the amount of cotton demanded be determined (1) when account is taken of the varying purchasing power of money; (2) when there is no variation in the purchasing power of money?

Let x_0 be the percentage change in the price of cotton, x_1 be the percentage change in the amount of cotton produced, and x_2 be the percentage change in the index number of general prices. We may then put our problem and its solution in this form:

(1) $x_0 = \phi(x_1, x_2)$, and we assume as a preliminary hypothesis that the form of $\phi(x_1, x_2)$ is linear so that we may write $x_0 = a_0 + a_1x_1 + a_2x_2$. According to the theory of multiple correlation, when the values of a_0, a_1, a_2 are so determined from the actual data as to make the correlation between the actual values of x_0 and the values of x_0 when forecast by the above formula a maximum, then

$$(x_0 - \bar{x}_0) = \frac{r_{01} - r_{02}r_{12}}{1 - r_{12}^2}\frac{\sigma_0}{\sigma_1}(x_1 - \bar{x}_1) + \frac{r_{02} - r_{01}r_{12}}{1 - r_{12}^2}\frac{\sigma_0}{\sigma_2}(x_2 - \bar{x}_2).$$

The statistical material necessary for the computation of the quantities indicated in these symbols is given in Table 33. When the actual computations are made and the numerical values are substituted in the above equation, we obtain as our forecasting formula

$$x_0 = -.97x_1 + 1.60x_2 + 7.11.$$

This formula enables us to predict the probable value of x_0 for given values of x_1, x_2; it enables us to say what

TABLE 33. — Data for the Quantitative Determination of the Law of Demand for Cotton

Year	Equivalent 500 Pound Bales, Gross Weight (Millions of Bales)	Price per Pound of Upland Cotton (Cents)	Bureau of Labor's Index of Prices of "All Commodities"	Percentage Change in the Amount Produced x_1	Percentage Change in the Price of Cotton x_0	Percentage in the Index of General Prices x_2
1889	7.47	11.5	115			
1890	8.56	8.6	113	+ 14.59	— 25.22	— 1.74
1	8.94	7.3	112	+ 4.44	— 15.12	— 0.88
2	6.66	8.4	106	— 25.50	+ 15.07	— 5.36
3	7.43	7.5	106	+ 11.56	— 10.71	0.00
4	10.03	5.9	96	+ 34.99	— 21.33	— 9.43
5	7.15	8.2	94	— 28.71	+ 38.98	— 2.08
6	8.52	7.3	90	+ 19.16	— 10.98	— 4.26
7	10.99	5.6	90	+ 28.99	— 23.29	0.00
8	11.44	4.9	93	+ 4.10	— 12.50	+ 3.33
9	9.35	7.6	102	— 18.27	+ 55.10	+ 9.68
1900	10.12	9.3	110	+ 8.24	+ 22.37	+ 7.84
1	9.51	8.1	108	— 6.03	— 12.90	— 1.82
2	10.63	8.2	113	+ 11.78	+ 1.23	+ 4.63
3	9.85	12.2	114	— 7.34	+ 48.78	+ 0.88
4	13.44	8.7	113	+ 36.45	— 28.69	— 0.88
5	10.58	10.9	116	— 21.28	+ 25.29	+ 2.65
6	13.27	10.0	122	+ 25.43	— 8.26	+ 5.17
7	11.11	11.5	130	— 16.28	+ 15.00	+ 6.56
8	13.24	9.2	123	+ 19.17	— 20.00	— 5.38
9	10.00	14.3	126	— 24.47	+ 55.43	+ 2.44
1910	11.61	14.7	132	+ 16.10	+ 2.80	+ 4.76
11	15.69	9.7	129	+ 35.14	— 34.01	— 2.27
12	13.70	12.0	134	— 12.68	+ 23.71	+ 3.88
13	14.16	13.1	135	+ 3.36	+ 9.17	+ 0.75

the probable change in the price of cotton will be when we know the probable changes in the production of cotton and in the level of general prices. Figure 12 traces for a period of twenty-five years the actual variations in the percentage changes in the price of

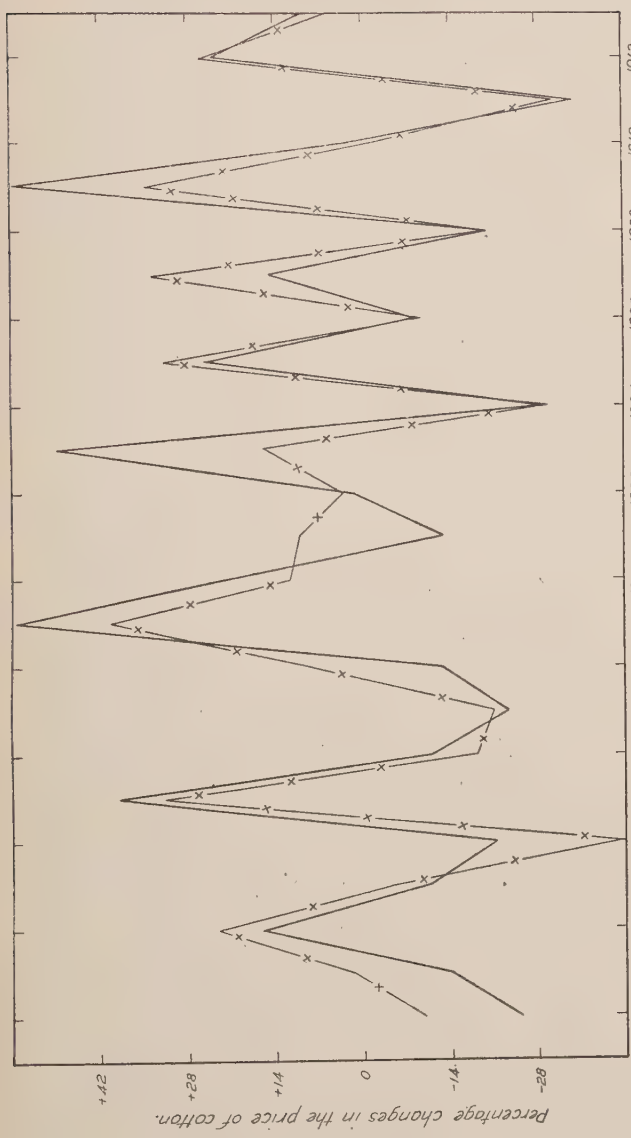

FIGURE 12. — The actual percentage changes in the prices of cotton, —, and the percentage changes as they are predicted, ×—×, by means of the dynamic law of demand for cotton, $x_0 = -.97x_1 + 1.60x_2 + 7.11$, where account is taken of the variations in the value of money.

cotton, together with the percentage changes as they are predicted by means of the formula $x_0 = -.97x_1 + 1.60x_2 + 7.11$.

(2) The degree of accuracy with which the above forecasting formula enables us to predict the changes in the price of cotton is measured by

$$S = \sigma_0\sqrt{1 - R^2}, \text{ where } R^2 = \frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{1 - r_{12}^2}.$$

The computations from the statistical data of Table 33 shows that

$$R = .859, \text{ and } S = 13.56.$$

This is a very high coefficient of correlation, and consequently the forecasting formula makes possible the prediction of the changes in the price of cotton with a relatively high degree of precision.

We have now a solution of the first part of our problem. We know how to forecast the changes in the price of cotton when account is taken of the changes in the amount demanded and of variations in the purchasing power of money. We know, besides, the degree of reliability with which our forecasts are made. We next enter upon the second part of our problem: What is the relation between the changes in the price of cotton and the changes in the amount demanded when there are no changes in the purchasing power of money?

(3) Since, in the forecasting formula $x_0 = -.97x_1 + 1.60x_2 + 7.11$, the variables x_0, x_1, x_2 are percentage changes, if we put $x_2 = 0$, we obtain the answer to the second part of our problem. The equation $x_0 = -.97x_1 + 7.11$ expresses the relation between the changes in

the price of cotton and the changes in the amount of cotton demanded *when the purchasing power of money remains constant.* Figure 13 traces the course of the actual changes in the price of cotton, and the changes as they would occur under the supposition that the level of general prices remains constant. The root-mean-square error of the forecasts by means of this formula is $S = 15.38$.

Furthermore, by the theory of partial correlation we know that when x_2 is constant — in this case when x_2 = zero — the coefficient measuring the relation between x_0 and x_1 is $\rho_{01} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1 - r_{02}^2}\sqrt{1 - r_{12}^2}}$, which, by computation from the statistical data of Table 33, gives $\rho_{01} = -.808$.

(4) If we collect our results bearing upon the relation of changes in the price of cotton, changes in the amount of cotton demanded, and changes in the purchasing power of money, we find that we have considered their interrelations under three different aspects:

(i) The relation between x_0 and x_1, when no attention is paid to the variable x_2, and x_0 is regarded as a simple function of x_1. This is the case of the dynamic law of demand in its simplest form. Here $r_{01} = -.819$; $S = 15.18$. The graph is given in Figure 11.

(ii) The relation between x_0 and both x_1 and x_2, where x_0 is regarded as a function of two variables. This is illustrative of the case of the dynamic law of demand in its complex form.

Here $R = .859$; $S = 13.56$. The graph is given in Figure 12.

(iii) The relation between x_0 and x_1, when $x_2 = 0$;

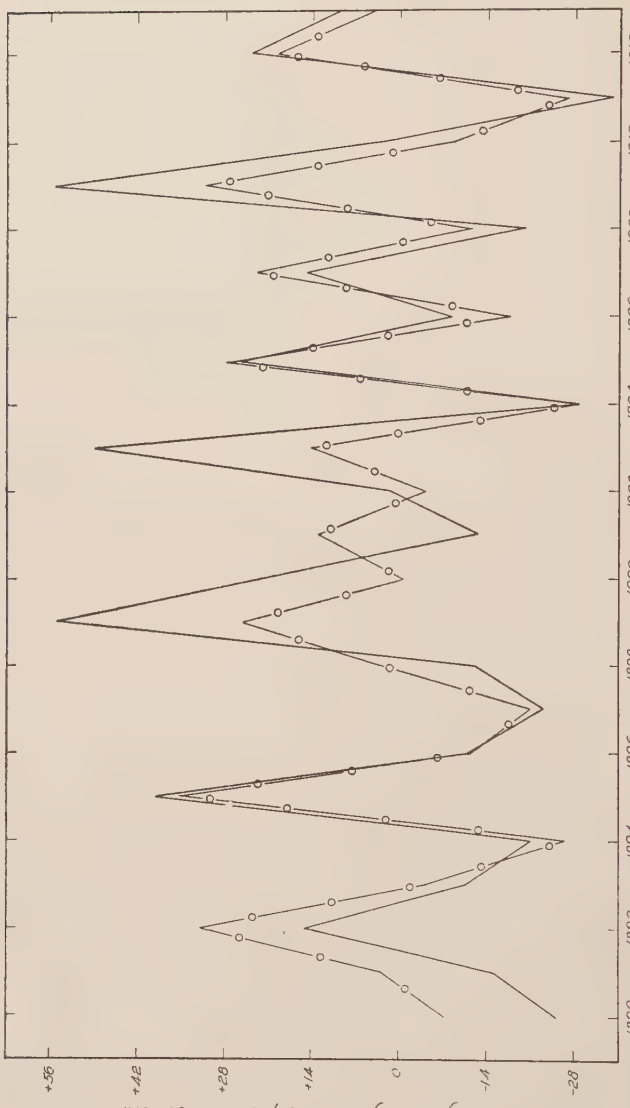

FIGURE 13. — The actual percentage changes in the prices of cotton, —, and the percentage changes as they are predicted, o—o, by means of the statical law of demand for cotton, $x_0 = -.97x_1 + 7.11$, where the purchasing power of money is supposed to remain constant.

that is to say the relation between the changes in the price of cotton and the changes in the amount of cotton demanded, when the purchasing power of money remains constant. This is illustrative of the static law of demand.

Here $\rho_{01} = -.808$; $S = 15.38$. The graph is given in Figure 13.

A consideration of these results will show how theoretical difficulties disappear before a practical solution. One of the discouraging aspects of deductive, mathematical economics is that when a complete theoretical formulation is given of the possible relations of factors in a particular problem, one despairs of ever arriving at a concrete solution because of the multiplicity of the interrelated variables. But the attempt to give statistical form to the equations expressing the interrelations of the variables shows that many of the hypothetical relations have no significance which needs to be regarded in the practical situation. When we write the law of demand in the form $x_0 = \phi(x_1, x_2, x_3, \ldots x_n)$, it is true, as we have pointed out, that we do not know the form of ϕ nor the types of the interrelations of $x_1, x_2, x_3, \ldots x_n$; but when we are confronted with the practical problem of forecasting x_0, we can find empirical functions that enable us to predict x_0 with a high degree of accuracy. Nor is this the only result of the practical solution. We find that since $r_{01} = -.819$ and $\rho_{01} = -.808$, there is really no difference in the closeness of the relation between x_0 and x_1 whether we completely ignore the variations in the purchasing power of money or regard the purchasing power of money

constant. And since $R = .859$, we learn that while the relation between x_0 and x_2 may be fairly high (in this case $r_{02} = .492$) still there is only a small advantage in accuracy of forecasting when we consider x_0 a function of the two variables x_1, x_2 instead of a simple linear function of x_1. If we were to regard $x_0 = \phi(x_1, x_2, x_3, \ldots x_n)$ and the correlations between x_0 and $x_3; \ldots;$ x_0 and x_n were small, little or nothing would be gained in accuracy of the forecast by considering these additional variables.[1]

The method which we have adopted in case of three variables is general in its character and may be applied to any number of variables. When $x_0 = \phi(x_1, x_2, x_3, \ldots x_n)$ and the variables are all percentage changes, it is possible not only to deal with the dynamic law of demand in all of its natural complexity, but also to ascertain the static law of demand giving the relation between x_0 and x_1 when all of the other variables are equal to zero.

The problem of ascertaining the statistical form of the law of demand receives by this method an adequate solution.

[1] "If A in part determines B, when we disregard other factors, and C in part determines B, when we disregard all else, and similarly D and E, it is argued that all these part-determinations can be added together and the sum will finally determine B. But the error made lies in the supposition that A, C, D, E, etc., are themselves *independent*. In the universe as we know it, all these factors are themselves to a greater or less extent associated or correlated, and in actual experience, but little effect is produced in lessening the variability of B, by introducing additional factors after we have taken the first few most highly associated phenomena." Pearson: *Grammar of Science*, 3rd edition, p. 172.

CHAPTER VI

CONCLUSIONS

The business of economic science, as distinguished from economic practice, is to discover the routine in economic affairs. It aims to separate out the elements of the routine, to ascertain their interrelations, and to use the knowledge of their connections to anticipate experience by forecasting from known changes the probabilities of correlated changes. The seal of the true science is the confirmation of the forecasts; its value is measured by the control it enables us to exercise over ourselves and our environment.

Economists theoretical and practical have grown impatient with any form of speculation that is not of immediate use. The present generation of theoretical economists expects an inquiry to be dynamic, to take account of the economic flux, to show a routine in change; otherwise, it is hypothetical, static, without significance in the affairs of daily life. The man of affairs must be convinced that an economic inquiry will either make directly for the common weal or else will reveal to him, in the pursuits of his daily life, a source of individual profit; otherwise, as far as he is concerned, the inquiry is academic, visionary, doctrinaire. The progress of the new type of economic theory is insured by the fact that it is profitable for practical men to give it their support.

Forecasting is the essential aim of both the economic scientist and the man of affairs. According to the most approved doctrine, economic profit has its origin in economic changes. Other forms of income — interest, wages, and rent — would exist in a purely stationary

state, but there would be no profit. The talent of the director of industry in the modern state consists in his capacity to foresee and to exploit economic changes, and his profit is proportionate to the accuracy with which his forecasts are made. The economic scientist is likewise concerned with changes. His talent consists in his capacity to separate the general from the accidental, to detect the routine in the multitudinous details. His success is proportionate to the simplicity and generality of the routine that he may discover and the accuracy with which he is able to foretell the size and direction of future changes.

To exemplify the simple laws of economic change, it appeared advisable to begin, not with a complex industrial state like England or the United States, but with a contemporary, progressive society in which the whole economic life is dependent upon a few fundamental interests. It seemed that no territory would afford a more promising field for such a quest than the Cotton Belt of the United States. Throughout a long period it has been recognized that in the vast area of the Cotton Belt which, with Russia excepted, equals in area a third of Europe, "*Cotton is King.*" And not only is cotton the leading staple of the South, but three-fourths of the world's production of this indispensable commodity is the yield of our Cotton Belt. Not only does the change in the price and yield of this commodity affect the local Cotton Belt, but, to the extent that cotton enters into international trade, its vicissitudes are reflected throughout the world.

Would it be possible to discover the routine in the

yield and price of cotton so that the knowledge might be used for purposes of forecasting?

For their information as to the condition and promise of the growing cotton crop, farmers, brokers, manufacturers, and merchants rely primarily upon the reports of the United States Department of Agriculture. To meet the public demand, the Department of Agriculture has instituted a wonderful statistical organization. By a connection with many thousands of correspondents, by field-agents, by special experts in crop estimates, by a Bureau of Statistics and a Crop-Reporting Board, information has been systematically gathered and tabulated, and for several decades monthly reports have been issued throughout the growth season of the crop. Extraordinary precautions have been taken to prevent any leakage of the precious information before it is given to the public.

What is the value of these reports? Since they are issued under the ægis of the Government they are assumed to be fairly accurate representations of the facts, and official authorities have, very naturally, lost no opportunity to point out the direct advantage to farmers of the expenditure of public funds for this particular purpose. Speculators have regarded the official documents as of value for their ends, and numerous rumors have circulated of bribes offered and bribes taken for advanced information as to the contents of the reports. But what is the value of these crop reports in the sense of their degree of accuracy as descriptions of actual facts and their measure of reliability as forecasts?

Five reports are issued during the growth season of

cotton and refer to the condition of the crop at the end of May, June, July, August, and September. An examination of these reports for a period extending over a quarter of a century, 1890–1914, shows:

(1) That the May report, covering the condition of the cotton crop in the whole country at the end of May, is so erroneous that any forecast from it is spurious. Any money that changes hands as a result of the report is the gain or loss of a simple gamble;

(2) That the June report as a basis of forecasts is better than the May report, but that its value for purposes of forecasting the yield per acre of cotton is negligible;

(3) That the remaining three reports — for July, August, and September — have real value, but the measurement of their degree of accuracy reveals the anomaly of the July report being as good as the report for August;

(4) That the official method of forecasting favors the farmers by giving an underestimate of the probable yield of cotton.

These are the concrete facts upon which the practical man in touch with actual affairs bases his economic conduct. Is it the part of a visionary to expect to obtain equally reliable forecasts of the cotton yield from the simple reports of the weather?

Lord Kelvin has told us that "when you can measure what you are speaking about and express it in numbers, you know something about it, but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind." By the help of statistical methods that rest

upon the theory of probability, it is possible to measure the precise degree of accuracy of any method of forecasting, and, consequently, it is possible to compare the relative accuracy of forecasts based upon official reports and forecasts that are derived from the records of accumulated rainfall and temperature in the states of the Cotton Belt. For purposes of comparison we have taken the four leading cotton states, which together produce 65 per cent of the entire crop. These four states are Texas, Georgia, Alabama, and South Carolina. Not only do these four states produce the greater part of the total cotton crop, but they represent the weather conditions throughout the whole Cotton Belt: Texas exemplifies the conditions in the extreme Southwest; Georgia and South Carolina, those at the other extreme on the Atlantic Coast; and Alabama typifies the conditions on the Gulf of Mexico. We shall consider, for these representative states, the results of comparing the forecasts from the condition of the growing crop, by the official method; and the forecasts from the changes in rainfall and temperature, by a method which we have fully described.

The comparison of methods will be made clear by examining first the results for the single state of Georgia. From calculations based upon data covering a quarter of a century, we find in case of Georgia:

(1) That for each of the five months of the growth season the forecast of the yield per acre of cotton which is based upon the weather data is decidedly better than the forecast from the condition of the crop, by means of the official formula;

(2) That for every month the forecasts from the

weather are better than the forecasts *a month later*, by the official method; or, more definitely, the forecasts from the accumulated weather at the end of May, June, July, and August are better than the forecasts by the official method at the end of June, July, August, and September;

(3) That when regard is paid to the probable errors of the coefficients measuring the accuracy of the forecasts, then, for every month, the forecasts from the weather are as good as the forecasts *two months later* by the official method. Or, more definitely, the forecast from the May weather is as good as the forecast by the official method at the end of July; the forecast from the joint effect of the May and June weather is as good as the forecast by the official method at the end of August, and the forecast from the accumulated weather at the end of July is as good as the forecast by the official method at the end of September.

We shall now extend our comparison to the results for the representative states, Texas, Georgia, Alabama, and South Carolina. As there are five monthly reports on the condition of the growing crop and we have taken four representative states, there are twenty cases in which the forecasts of the yield per acre of cotton may be compared:

(1) In 17 out of 20 cases the forecasts from the weather are more accurate than the forecasts from the condition of the crop, by the official method;

(2) For all of the representative states the forecasts by the official method from the May condition of the crop are worthless. By contrast, all of the fore-

casts from the May weather have value. The forecasts from the weather for Georgia and South Carolina are, at the end of May, better than the forecasts by the official method at the end of June, and about as good as those at the end of July; and the forecast from the May weather in Alabama is about as good as the forecast by the official method at the end of September. The value of the forecast from the May weather in Texas is negligible;

(3) For three out of the four representative states the forecasts from the June condition of the crop, by means of the official method, are worthless. But in all three cases the forecasts from the accumulated weather at the end of June are better than the forecasts by the official method at the end of July;

(4) For all of the states except Texas the forecasts from the weather give, for each month, more accurate predictions than can be obtained by the official method from the condition of the crop *one month later*. The forecasts from the accumulated weather at the end of May, June, July, and August are better than the forecasts by the official method at the end of June, July, August, and September;

(5) For all of the states except Texas the forecasts from the accumulated weather at the end of May, June, and July are about as good as can be obtained by the official method from the condition of the crop *two months later*, at the end, respectively, of July, August, and September.

As the routine of measurable dependence of yield upon the weather is due to the presence of natural

causes, it might easily be inferred that when we move to strictly social facts, no such routine will be found. It could be argued that the price of cotton results from "the law of supply and demand"; the supply may be predictable because it is primarily dependent upon natural causes; but the demand is a social fact and is the resultant of many individual choices each of which, in its turn, is dependent upon many variable factors. By such *a priori* reasoning it would be easy to conclude that it is futile to attempt to find a predictable routine in the dependence of the price of cotton upon the size of the crop.

But again we are reminded of Lord Kelvin's statement that "when you can measure what you are speaking about and express it in numbers, you know something about it, but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind." Our researches have shown that there is a dynamic law of demand — a law that connects the price of cotton with the size of the crop — and that the knowledge of the law would have made possible the prediction of the price of cotton from 1890 to 1914 with a degree of accuracy higher than that attained by the formula of the Department of Agriculture in the annual prediction of the size of the crop at the end of September. Upon the appearance of the monthly cotton reports, great sums of money exchange hands because of the light they are supposed to throw upon the probable size of the crop. The most reliable report is, of course, the one nearest the harvest, but the accuracy of the forecasts of yield that are based upon this report is less than the accuracy with which

the price of cotton can be predicted from the size of the crop, by means of the law of demand.

Both the yield and the price of cotton, therefore, are so much a matter of routine that they admit of prediction with a high degree of precision.

In laying the foundation of the modern type of Economic Theory, Jevons foresaw the necessity of the simultaneous development of its deductive and its inductive phases:

"I know not when we shall have a perfect system of statistics, but the want of it is the only insuperable obstacle in the way of making Economics an exact science. In the absence of complete statistics, the science will not be less mathematical, though it will be immensely less useful than if it were, comparatively speaking, exact. A correct theory is the first step towards improvement, by showing what we need and what we might accomplish."

"The deductive science of Economics must be verified and rendered useful by the purely empirical science of statistics. Theory must be invested with the reality and life of fact." [1]

In the opinion of Professor Marshall the pressing need of economic science at the present time is "the quantitative determination of the relative strength of different economic forces." [2] And Professor Pareto, in

[1] Jevons: *Theory of Political Economy*, 3rd edition, pp. 12, 22. Cf. Jevons: *Principles of Science*, chapter xxii, end of the section on "Illustration of Empirical Quantitative Laws."

[2] Cf. the address of Professor W. J. Ashley as President of Section F of the British Association for the Advancement of Science. *Report of the British Association for the Advancement of Science*, 1907, p. 591.

like spirit, points out the conditions of the further development of our science:

"The progress of Political Economy in the future will depend in great part upon the investigation of empirical laws, derived from statistics, which will then be compared with known theoretical laws, or will suggest the derivation from them of new laws." [1]

The idea that I should like to emphasize is that because of the recent development of statistical theory and the improvement in the collection of statistical data, we are now able to meet the needs so clearly described by the masters of the science.

The great advance in the methodology of deductive economics, after Cournot's epoch-making work, was initiated by Léon Walras in his use of simultaneous equations for the purpose of completely surveying the interrelated factors in the problems of exchange, production, and distribution. It was necessary in his work and in the work of his successors to begin with a simple, hypothetical construction and to approach the concrete problem by the introduction of an increasing number of complicating factors. The equations expressing the relations between the variables were, of necessity, arbitrary, but the device made possible the envisaging of all the elements in the problem, and suggested the types of their interrelations. But economic theory has now reached the stage where, according to Professor Marshall, there is need of a "quantitative determination of the relative strength of the different economic

[1] *Giornale degli Economisti,* Maggio, 1907, p. 366. "Il progresso dell' Economia politica dipenderà pel futuro in gran parte dalla ricerca di leggi empiriche, ricavate dalla statistica, e che si paragoneranno poi colle leggi teoriche note, o che ne faranno conoscere di nuove."

forces"; and, according to Professor Pareto, empirical laws must be derived from statistics for the double purpose of comparing them with known theoretical laws and of gaining bases for new theoretical developments.

The statistical theory of multiple correlation is perfectly adapted to these demands. No matter what may be the number of factors in the economic problem, it is specially fitted to make a "quantitative determination" of their relative strength; and no matter how complex the functional relations between the variables, it can derive "empirical laws" which, by successive approximations, will describe the real relations with increasing accuracy. The mathematical method of deductive economics gives a *coup d'œil* of the factors in the problem; the statistical method of multiple correlation affixes their relative value and reveals the laws of their association. The mathematical method begins with an ultra-hypothetical construction and then, by successive complications, approaches a theoretical description of the concrete goal. The method of multiple correlation reverses the process: It begins with concrete reality in all of its natural complexity and proceeds to segregate the important factors, to measure their relative strength, and to ascertain the laws according to which they produce their joint effect. When the method of multiple correlation is thus applied to economic data it invests the findings of deductive economics with "the reality and life of fact"; it is the Statistical Complement of Deductive Economics.

THE following pages contain advertisements of Macmillan books by the same author.

Economic Cycles: Their Law and Cause

By HENRY LUDWELL MOORE

Professor of Political Economy in Columbia University

8vo, $2.00

Extract from the Introduction: "There is a considerable unanimity of opinion among experts that, from the purely economic point of view, the most general and characteristic phenomenon of a changing society is the ebb and flow of economic life, the alternation of energetic, buoyant activity with a spiritless, depressed and uncertain drifting. . . . What is the cause of this alternation of periods of activity and depression? What is its law? These are the fundamental problems of economic dynamics the solution of which is offered in this Essay."

Comments of Specialists

Moore's book is so important that it is sure to be widely criticized. . . . Yet so far as the fundamental conclusions are concerned the book is so firmly grounded on a vast body of facts that its main line of argument seems unassailable. . . . Moore has gone much further than his predecessors and has removed his subject from the realm of probability to that of almost absolute certainty. Hereafter there can be little question that apart from such influences as the depreciation in gold, or great calamities like the war, the general trend of economic conditions in this country is closely dependent upon cyclical variations in the weather." — Ellsworth Huntington, in the *Geographical Review*.

In reply to the question: "What are the two best books you have read recently," President Butler named, as one of the two books, Professor Moore's *Economic Cycles* because of its being "an

original and very stimulating study in economic theory with quick applications to practical business affairs." — Nicholas Murray Butler, in the *New York World.*

"Professor Moore is known among scholars as one of the keenest and most cautious of investigators. . . . His novel methods of investigation constitute an additional claim upon our interest; the problem of the crisis has never yet been approached in precisely this way." — Alvin S. Johnson, in the *New Republic.*

"This book indicates a method of utilizing (economic) data . . . that is worthy of the highest commendation." — Allen Hazen, in the *Engineering News.*

"If the promise of Professor Moore's convincing Essay is fulfilled, economics will become an approximately exact science. . . . If progress is made in the direction of such a goal as a result of this work, it will be the economic contribution of a century, and will usher in a new scientific epoch." — Roy G. Blakey, in the *Times Annalist.*

"The agricultural theory of cycles has found a new and brilliant exponent in Professor Henry L. Moore." — Wesley Clair Mitchell, in the *American Yearbook.*

"If his methods stand the test of experience, and can be widely adopted, the field of business may be revolutionized so far as it concerns the enterpriser because the measuring of the force of underlying, fundamental conditions will become approximately accurate and the function of the enterpriser will thereby be reduced." Magazine published by *Alexander Hamilton Institute.*

"L'auteur a mis à son service des procédés mathématiques et statistiques raffinés et élégants . . . celui-ci a écrit un livre brillant." — Umberto Ricci, in *Scientia.*

THE MACMILLAN COMPANY
Publishers **64–66 Fifth Avenue** **New York**

Laws of Wages

An Essay In Statistical Economics

By HENRY LUDWELL MOORE

Professor of Political Economy in Columbia University

8vo, $1.60

Extract from the Introduction: "In the following chapters I have endeavored to use the newer statistical methods and the more recent economic theory to extract, from data relating to wages, either new truth or else truth in such new form as will admit of its being brought into fruitful relations with the generalizations of economic science."

Comments of Specialists

"Professor Moore brings to his task a wide acquaintance with the most difficult parts of the literature of economics and statistics, a full appreciation of its large problems, a judicial spirit and a dignified style." — F. W. Taussig, in the *Quarterly Journal of Economics.*

"Statistics of the ordinary official kind have often served to support the arguments of political economists. But this is the first time, we believe, that the higher statistics, which are founded on the Calculus of Probabilities, have been used on a large scale as a buttress of economic theory." — F. Y. Edgeworth, in the *Economic Journal.*

"Professor Moore has broken new ground in a most interesting field, and while we may differ from him in the weight to be attached to this or that result or the interpretation to be placed on some

observed coefficient, we may offer cordial congratulations on the work as a whole." — G. Y. Yule, in the *Journal of the Royal Statistical Society.*

"Die Fruchtbarkeit der verwendeten Methode scheint mir durch diese Untersuchungen zweifellos erwiesen, ebenso wie die Erreichbarkeit des Ziels, die Theorie ganz dicht an die Zahlenausdrücke der wirtschaftlichen Tatsachen heranzubringen. Und das ist eine Tat, zu der der Autor nur zu beglückwünschen ist. . . . Hat das Buch auch auf der Hand liegende Fehler — in der Zukunft wird man sich seiner als der ersten klaren, einfachen und zielbewussten Darlegung und Exemplifizierung der Anwendung der 'höheren Statistik' auf ökonomische Probleme dankbar erinnern." — Joseph Schumpeter, in the *Archiv für Sozialwissenschaft und Sozialpolitik.*

"Non seulement il nous enseigne l'emploi d'une méthode qui dans de certaines limites peut être très féconde. Mais encore son habileté personnelle dans le maniement de cette méthode est très réelle. Il sait scruter les statistiques d'une façon fort pénétrante et exposer les résultats de ses recherches avec beaucoup d'élégance. Le lecteur français en particulier, appréciera l'ingéniosité avec laquelle il tire des statistiques françaises des inductions souvent nouvelles et justes." — Albert Aftalion, in the *Revue d'histoire des doctrines économiques.*

"Alcuni dei risultati ottenuti dall'autore, sono nuovi e suggestivi e da essi molte conclusioni si possono trarre (cui l'autore accenna nel capitolo finale della sua opera) sia rispetto alle teorie del salario che rispetto alla politica sociale. Il libro è insomma, ripetiamo, un contributo molto importante all'investigazione scientifica dei fenomeni economici e vorremmo che esso stimolasse parecchi altri studiosi a fare per altre industrie o per altri paesi, recerche analoghe." — Constantino Bresciani Turroni, in the *Giornale degli Economisti.*

THE MACMILLAN COMPANY
Publishers **64–66 Fifth Avenue** **New York**

Date Due

Library Bureau Cat. No. 1137

Dup BL

HD9074
.M7 179544

HD9074
.M7 179544

AUTHOR Moore, H.L.

TITLE Forecasting the yield and the price of cotton

DATE DUE	BORROWER'S NAME

ST. OLAF COLLEGE
HD9074 .M7
Moore, Henry Lu - Forecasting the yield and the pr
3 0111 00189 3329